THE BELL SYSTEM

Technical Journal

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

VOLUME XL MARCH 1961 NUMBER 2

COPYRIGHT 1961 AMERICAN TELEPHONE AND TELEGRAPH COMPANY

THE BELL SYSTEM TECHNICAL JOURNAL

ADVISORY BOARD

H. I. ROMNES, *President, Western Electric Company*

J. B. FISK, *President, Bell Telephone Laboratories*

E. J. MCNEELY, *Executive Vice President, American Telephone and Telegraph Company*

EDITORIAL COMMITTEE

A. C. DICKIESON, *Chairman*

S. E. BRILLHART
A. J. BUSCH
L. R. COOK
R. L. DIETZOLD
J. H. FELKER
K. E. GOULD
G. GRISWOLD, JR.
J. R. PIERCE
M. SPARKS
W. O. TURNER

EDITORIAL STAFF

G. E. SCHINDLER, JR., *Editor*
R. M. FOSTER, JR., *Assistant Editor*
C. POLOGE, *Production Editor*
J. T. MYSAK, *Technical Illustrations*
T. N. POPE, *Circulation Manager*

THE BELL SYSTEM TECHNICAL JOURNAL is published six times a year by the American Telephone and Telegraph Company, 195 Broadway, New York 7, N. Y. F. R. Kappel, President; S. Whitney Landon, Secretary; L. Chester May, Treasurer. Subscriptions are accepted at $5.00 per year. Single copies $1.25 each. Foreign postage is $1.08 per year or 18 cents per copy. Printed in U.S.A.

THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XL MARCH 1961 NUMBER 2

Copyright 1961, American Telephone and Telegraph Company

Pulse Transmission by AM, FM and PM in the Presence of Phase Distortion

By E. D. SUNDE

(Manuscript received October 13, 1960)

In pulse transmission systems, pulses modulated in various ways to carry information may be transmitted by amplitude, phase or frequency modulation of a carrier, and with each type of modulation various methods of detection are possible. An important consideration in many applications is the performance of various modulation and detection methods in the presence of phase distortion or equivalent envelope delay distortion, which may be appreciable in certain transmission facilities. The principal purpose of this presentation is a theoretical evaluation of transmission impairments resulting from certain representative types of delay distortion. These transmission impairments are reflected in the need for increased signal-to-noise ratio at the detector input to compensate for the effect of delay distortion.

The performance in pulse transmission by various carrier modulation and detection methods can be formulated in terms of a basic function common to all, known as the carrier pulse transmission characteristic, which is related by a Fourier integral to the amplitude and phase characteristics of the channel. Numerical values are given here for the carrier pulse transmission characteristic with linear and quadratic delay distortion, together with the maximum transmission impairments caused by these fairly representative forms of delay distortion with various methods of carrier modulation and signal detection. These include amplitude modulation with envelope and with synchronous detection, two-phase and four-phase modulation with synchronous detection and with differential phase detection and binary frequency modulation.

In determining the effect of delay distortion, a raised cosine amplitude spectrum of the pulses at the detector input has been assumed in all cases, together with the minimum pulse interval permitted with this spectrum and ideal implementation of each modulation and detection method. Furthermore, optimum adjustments from the standpoint of slicing levels and sampling instants at the detector output are assumed for each particular case of delay distortion. These idealizations insure that only the effect of delay distortion is evaluated and considered in comparing modulation methods, and that this effect is minimized by appropriate system adjustments.

TABLE OF CONTENTS

I. INTRODUCTION

Binary pulse transmission by various methods of carrier modulation has been dealt with elsewhere on the premise of ideal amplitude and phase characteristics of the carrier channels.[1] An important consideration in many applications is the performance in the presence of phase distortion or equivalent envelope delay distortion, which may be appreciable in certain transmission facilities. An ideal amplitude spectrum of received pulses can be approached with the aid of appropriate terminal filters with gradual cutoffs, such that the associated phase characteristic is virtually linear. Nevertheless, pronounced phase distortion may be encountered in pulse transmission over channels with sharp cutoffs outside the pulse spectrum band, as in frequency division carrier system channels designed primarily for voice transmission.

The principal purpose of the present analysis is a theoretical evaluation of transmission impairments resulting from certain representative types of delay distortion in pulse transmission by various methods of carrier modulation and signal detection. These transmission impairments are reflected in the need for increased signal-to-noise ratio at the detector input to compensate for the effect of delay distortion.

The performance in pulse transmission by various carrier modulation and detection methods can be related to a basic function known as the carrier pulse transmission characteristic. This basic function gives the shape of a single carrier pulse at the channel output, i.e., the detector input, under ideal conditions or in the presence of the particular kind of transmission distortion under consideration. From this basic function can be determined the envelopes of carrier pulse trains at the detector input, together with the phase of the carrier within the envelope. The shape of demodulated pulse trains with various methods of carrier modulation and detection can, in turn, be determined for various combinations of transmitted pulses, together with the maximum transmission impairment from a specified type of channel imperfection, such as delay distortion dealt with here.

The carrier pulse transmission characteristic is related by a Fourier integral to the amplitude and phase characteristic of the channel. It has been determined elsewhere[2] for pulses with a raised cosine spectrum and cosine variation in transmission delay over the channel band, and for pulses with a gaussian spectrum with linear variation in delay. A cosine variation in delay is approximated in some transmission facilities and has certain advantages from the standpoint of analysis, both as regards numerical evaluation and interpretation in terms of pulse echoes.

A somewhat similar form of delay distortion that affords a satisfactory

approximation in many cases is quadratic (or parabolic) delay distortion. Quadratic delay distortion is in theory approached near midband of a flat bandpass channel with sharp cutoffs, such as a carrier system voice channel, and usually affords a satisfactory approximation over the more important part of the transmission band of such channels. Linear delay distortion is approximated when a bandpass channel with gradual cutoffs is established to one side of midband of a flat bandpass channel with sharp cutoffs. These and other types of delay distortion do not lend themselves to convenient analytical evaluation of the Fourier integrals for the pulse transmission characteristic. However, at present, these integrals can be accurately evaluated by numerical integration with the aid of digital computers for any specified pulse spectrum and phase distortion.

Numerical values are given here for the carrier pulse transmission characteristics with linear and quadratic delay distortion, together with the maximum transmission impairments caused by these limiting and fairly representative forms of delay distortion with various methods of carrier modulation and signal detection. These include amplitude modulation with envelope and with synchronous detection, two-phase and four-phase modulation with synchronous detection and with differential phase detection and binary frequency modulation. In determining the effect of delay distortion, a raised cosine amplitude spectrum of the pulses at the detector input has been assumed in all cases, together with the minimum pulse interval permitted with this spectrum, ideal implementation of each modulation and detection method and optimum design from the standpoint of slicing levels and sampling instants at the detector output. These idealizations insure that only the effect of delay distortion is evaluated and considered in comparing modulation methods, a condition that is difficult to realize with experimental rather than analytical comparisons.

As mentioned above, the present analysis involves a basic function common to all modulation methods, which in general would be determined with the aid of digital computers. This approach has certain advantages in comparison of modulation methods and from the standpoint of optimum system design over direct computer simulation of each modulation method. The latter direct approach may be preferable for any specified modulation method and type of transmission impairment and has been used in connection with a binary double-sideband AM system with envelope detection, for cosine and sine variations in transmission delay over the channel band and for combinations of these.[3] The transmission impairments caused by linear and quadratic delay distortion, and combinations thereof, have been determined experi-

mentally for a binary vestigial-sideband amplitude modulation data transmission system employing envelope detection.[4]

The present analysis is concerned with certain "coarse structure" variations in transmission delay that ordinarily predominate over smaller "fine structure" variations, except in transmission facilities where elaborate phase equalization is used. Transmission impairments from small irregular fine structure gain and phase deviations over the channel band can be evaluated by methods discussed elsewhere[2] and are not considered here.

II. CARRIER PULSE TRAINS AND MODULATION METHODS

2.1 *General*

In carrier pulse modulation systems the pulse trains at the transmitting end modulate a carrier in amplitude, phase or frequency. In AM the demodulated signal depends on the envelope of the received carrier pulse train at sampling instants, in PM on the phase of the carrier within the envelope and in FM on the time derivative of the phase at sampling instants. To determine the performance of these various methods in the presence of transmission distortion it is necessary to formulate the received carrier pulse trains.

The received carrier pulse trains at the channel output, i.e., the detector input, can in all cases be formulated in terms of the carrier pulse transmission characteristic, that is, the received carrier pulse in response to a single transmitted pulse. This pulse transmission characteristic is related to the shape of the modulating pulses at the transmitting end, and to the amplitude and phase characteristic of the channel, by a Fourier integral, as discussed and illustrated for special cases in the Appendix. The general formulation of the pulse trains at the detector input and the resultant demodulated pulse trains with various methods of carrier modulation and signal detection is dealt with in the following sections.

2.2 *Carrier Pulse Transmission Characteristics*

It will be assumed that a carrier pulse of rectangular or other suitable envelope is applied at the transmitting end of a bandpass channel. The received pulse with carrier frequency ω_c can then be written in the general form [Ref. 2, Equation (2.09)]

$$P_c(t) = \cos(\omega_c t - \psi_c) R_c(t) + \sin(\omega_c t - \psi_c) Q_c(t) \tag{1}$$

$$= \cos[\omega_c t - \psi_c - \varphi_c(t)] \bar{P}_c(t), \tag{2}$$

where

$$\bar{P}_c(t) = [R_c^2(t) + Q_c^2(t)]^{\frac{1}{2}}, \tag{3}$$

$$\varphi_c(t) = \tan^{-1}[Q_c(t)/R_c(t)], \tag{4}$$

$$R_c(t) = \bar{P}_c(t) \cos \varphi_c(t), \tag{5}$$

$$Q_c(t) = \bar{P}_c(t) \sin \varphi_c(t). \tag{6}$$

In the above relations R_c and Q_c are the envelopes of the in-phase and quadrature components of the received carrier pulse and $\bar{P}_c$ the resultant envelope. The time t is taken with respect to a conveniently chosen origin, for example the midpoint of a pulse interval or the instant at which R_c or $\bar{P}_c$ reaches a maximum value.

With a carrier frequency ω_0 rather than ω_c, relation (1) is modified into

$$P_0(t) = \cos(\omega_0 t - \psi_0)R_0(t) + \sin(\omega_0 t - \psi_0)Q_0(t) \tag{7}$$

and relations (2) through (6) are correspondingly modified by replacing c by the subscript 0.

When the carrier frequency is changed from ω_0 to ω_c the spectrum of a received pulse will change, provided the transmission-frequency characteristic of the channel remains fixed, except in the limiting case of a carrier pulse of infinitesimal duration having a flat spectrum. However, by appropriate modification of the transmission-frequency characteristic the amplitude spectrum of a pulse at the channel output, i.e., the detector input, can be made the same regardless of the carrier frequency. On the latter premise of equal amplitude spectra at carrier frequencies ω_0 and ω_c, the following relations apply [Ref. 2, Equation (2.18)]:

$$\begin{aligned} R_c(t) &= \cos[\varphi_0(t) + \omega_y t - \psi_y]\bar{P}_0(t) \\ &= \cos(\omega_y t - \psi_y)R_0(t) - \sin(\omega_y t - \psi_y)Q_0(t), \end{aligned} \tag{8}$$

$$\begin{aligned} Q_c(t) &= \sin[\varphi_0(t) + \omega_y t - \psi_y]\bar{P}_0(t) \\ &= \cos(\omega_y t - \psi_y)Q_0(t) + \sin(\omega_y t - \psi_y)R_0(t), \end{aligned} \tag{9}$$

where

$$\varphi_0(t) = \tan^{-1}[Q_0(t)/R_0(t)],$$

$$\omega_y = \omega_c - \omega_0,$$

$$\psi_y = \psi_c - \psi_0.$$

Relations (8) and (9) apply when $R_c(t)$ and $Q_c(t)$ are referred to the

carrier phase ψ_c in (1) rather than the carrier phase ψ_0 in (7). If a carrier phase ψ_0 is used as reference, $\psi_y = 0$ in (8) and (9), and

$$R_{c.0} = \cos \omega_y t R_0(t) - \sin \omega_y t Q_0(t), \tag{10}$$

$$Q_{c.0} = \cos \omega_y t Q_0(t) + \sin \omega_y t R_0(t). \tag{11}$$

With (8) and (9) in (3), or (10) and (11) in (3):

$$\bar{P}_c(t) = \bar{P}_0(t) = [R_0^2(t) + Q_0^2(t)]^{\frac{1}{2}}. \tag{12}$$

The resultant envelope of a single pulse is thus the same regardless of carrier frequency and phase, on the premise of a fixed pulse spectrum at the channel output as assumed above.

2.3 *Pulse Trains at Detector Input*

Let carrier pulses be transmitted at intervals T, and let t be the time from the midpoint of a selected interval. The following designation will be introduced for convenience

$$\begin{aligned} R_c(t + nT) &= R_c\left[T\left(\frac{t}{T} + n\right)\right] = R_c(x + n), \\ Q_c(t + nT) &= Q_c\left[T\left(\frac{t}{T} + n\right)\right] = Q_c(x + n), \end{aligned} \tag{13}$$

where n is the time expressed in an integral number of pulse intervals of duration T and x the time in a fraction of a pulse interval.

Let $a(-n)$ and $\psi_c(-n)$ be the amplitude and phase of the carrier pulse transmitted in the nth interval prior to the interval 0 under consideration, and $a(n)$, $\psi_c(n)$ the corresponding quantities for the nth subsequent interval. The received pulse train in the interval $-T/2 < t < T/2$ is then

$$\begin{aligned} W_0(x) &= \sum a(n) \cos [\omega_c t - \psi_c(n)] R_c(x - n) \\ &\quad + \sum a(n) \sin [\omega_c t - \psi_c(n)] Q_c(x - n) \end{aligned} \tag{14}$$

$$= \sum a(n) \cos [\omega_c t - \psi_c(n) - \varphi_c(x - n)] \bar{P}_c(x - n), \tag{15}$$

where the summations are between $n = -\infty$ and $n = \infty$.

During the next interval, T to $2T$, the received wave is obtained by replacing $a(n)$ and $\psi_c(n)$ by $a(n + 1)$ and $\psi_c(n + 1)$ and is thus

$$W_1(x) = \sum a(n + 1) \cos [\omega_0 t - \psi_c(n + 1) - \varphi_c(x - n)] \bar{P}_c(x - n) \tag{16}$$

where t and x refer to midpoint of interval 1.

In pulse modulation systems as considered herein it is assumed that the modulating pulses are rectangular in shape and of duration equal to the pulse interval. For equal phases $\psi_c(n) = \psi_c$ of all the modulating pulses, (14) then becomes

$$W_0 = \cos(\omega_c t - \psi_c) \sum a(n) R_c(x - n) + \sin(\omega_c t - \psi_c) \sum a(n) Q_c(x - n). \tag{17}$$

When $a(n) = a$ is a constant the input is a continuous carrier, so that evaluation of (17) will give

$$W_0 = a A_c \cos(\omega_c t - \psi_c), \tag{18}$$

where A_c is the amplitude of the transmission-frequency characteristic of the channel at $\omega = \omega_c$ and it is assumed in the determination of R_c and Q_c that the phase characteristic is zero at $\omega = \omega_c$. That is, a constant transmission delay is ignored, which is permissible without loss of generality.

When $R_c(t)$ and $Q_c(t)$ are determined from the channel transmission-frequency characteristic by the usual Fourier integral relations, in the form represented by (159) through (163) of the Appendix, the following relations apply for rectangular modulating pulses of duration T equal to the pulse interval:

$$A_c = \sum_{n=-\infty}^{\infty} R_c(x - n), \tag{19}$$

$$0 = \sum_{n=-\infty}^{\infty} Q_c(x - n). \tag{20}$$

2.4 *Amplitude Modulation*

In AM systems $\psi_c(n) = \psi_c =$ constant and (14) becomes

$$W_0(x) = \cos(\omega_c t - \psi_c) \sum a(n) R_c(x - n) + \sin(\omega_c t - \psi_c) \sum a(n) Q_c(x - n). \tag{21}$$

With synchronous detection, also referred to as homodyne and coherent detection, the received wave is applied to a product demodulator together with a demodulating wave $\cos(\omega_0 t - \psi_c)$. After elimination of higher frequency demodulation products by low-pass filtering the demodulated baseband output becomes, when a factor of one-half is omitted for convenience

$$U_0(x) = \sum a(n) R_c(x - n). \tag{22}$$

If ω_s is the bandwidth of the modulating signal, the high-frequency output of the product demodulator will have a lowest frequency $2\omega_c - \omega_s$, which can be separated from the modulating wave by low-pass filtering provided $2\omega_c - \omega_s \geqq \omega_s$, or if $\omega_c \geqq \omega_s$.

At sampling instants $x = 0$, the desired signal is $a(0)R(0)$ and the remaining terms in (22) represent intersymbol interference in systems where $R_c(n) \neq 0$ for $n = \pm 1, \pm 2$, etc.

Owing to elimination of the quadrature components, synchronous detection is simpler from the standpoint of analysis than envelope detection, in which the demodulated signal depends on the envelope of the received wave (21) as given by

$$\bar{W}_0(x) = \{[\sum a(n)R_c(x - n)]^2 + [\sum a(n)Q_c(x - n)]^2\}^{\frac{1}{2}}. \quad (23)$$

The desired signal at sampling instants $x = 0$ is $a(0)[R_c^2(0) + Q_c^2(0)]^{\frac{1}{2}}$ and the remaining terms in (23) represent intersymbol interference.

2.5 *Phase Modulation with Synchronous Detection*

In phase modulation systems the amplitude $a(n) = a = \text{constant}$ and the phase $\psi_c(n)$ is varied from one pulse interval to the next. The received wave (15) then becomes

$$W_0(x) = \sum \cos [\omega_c t - \psi_c(n) - \varphi_c(x - n)]\bar{P}_c(x - n). \quad (24)$$

In a multiphase system, the received wave is in general applied to several product demodulators together with a demodulating wave $\cos(\omega_c t - \psi)$. In the particular case of two-phase modulation a single demodulator suffices, and the demodulator output after elimination of high-frequency demodulation products by low-pass filtering and omitting a factor of one-half is of the general form

$$U_0(x) = \sum \cos [\psi_c(n) - \psi + \varphi_c(x - n)]\bar{P}_c(x - n). \quad (25)$$

At the sampling instants

$$U_0(0) = \sum \cos [\psi_c(n) - \psi + \varphi_c(-n)]\bar{P}_c(-n), \quad (26)$$

where as before the summation is between $n = -\infty$ and $n = \infty$.

The desired signal is represented by the term for $n = 0$ and is

$$\begin{aligned} U_0(0) &= \cos [\psi_c(0) - \theta]\bar{P}_c(0) \\ &= [\cos \theta \cos \psi_c(0) + \sin \theta \sin \psi_c(0)]\bar{P}_c(0), \end{aligned} \quad (27)$$

where

$$\theta = \psi - \varphi_c(0). \quad (28)$$

When the phase ψ of the demodulating wave is so chosen that $\theta = 0$, and if $\psi_c(0) = 0$ or π as in two-phase modulation, then $U_0(0) = \pm\bar{P}_c(0)$.

In four-phase modulation two product demodulators are required, with the demodulating waves displaced 90° in phase. The output of the second demodulator is then, in place of (26),

$$V_0(0) = \sum \sin [\psi_c(n) - \psi + \varphi_c(x - n)]\bar{P}_c(-n) \tag{29}$$

and the desired output at sampling instants is, in place of (27),

$$\begin{aligned} V_0(0) &= \sin [\psi_c(0) - \theta]\bar{P}_c(0) \\ &= [\cos\theta \sin \psi_c(0) - \sin\theta \cos \psi_c(0)]\bar{P}_c(0). \end{aligned} \tag{30}$$

The preferable choice of the phase of the demodulating wave in the above relations may depend on certain considerations in the implementation of modulators and demodulators. In Table I are given the four possible combined outputs as determined by the carrier phase $\psi_c(0)$ for the particular cases $\theta = 0$ and $\theta = \pi/4$. For convenience the outputs for $\theta = \pi/4$ are normalized to unit amplitude, the actual amplitudes being $\pm\frac{1}{2}\sqrt{2}$.

It will be noted that with $\theta = 0$ the output U_0^0 determines whether one carrier is modulated in phase by $\psi_c = 0$ or π, while the output V_0^0 determines whether the quadrature carrier is modulated in phase by $\psi_c = 0$ or π. The two carriers can thus be modulated and demodulated independently, without the need for circuitry to convert the two demodulator output to carrier phase, as would be required with $\theta = \pi/4$. With differential phase detection, to be discussed in the next section, such converters would be required both with $\theta = 0$ and $\theta = \pi/4$. In this case $\theta = \pi/4$ may be preferable for the reason that only two states $(-1,1)$ are possible for each demodulator, rather than three states $(-1,0,1)$ with $\theta = 0$.

TABLE I — DEMODULATOR OUTPUTS U_0^0 AND V_0^0 IN FOUR-PHASE SYSTEMS AS DETERMINED BY CARRIER PHASE $\psi_c(0)$ FOR DEMODULATING WAVES WITH PHASES $\theta = 0$ AND $\pi/4$

$\psi_c(0)$	$\theta = 0$		$\theta = \pi/4$	
	U_0^0	V_0^0	U_0^0	V_0^0
0	1	0	1	−1
$\pi/2$	0	1	1	1
π	−1	0	−1	1
$3\pi/2$	0	−1	−1	−1

2.6 *PM with Differential Phase Detection*

An alternative method of demodulation that will be considered in connection with phase modulation is differential phase detection. With this method $W_0(x)$ as given by (15) is applied to one pair of terminals of a product demodulator, and $W_1(x)$ as given by (16) to the other pair with a suitable phase shift θ. The demodulator output is then, with $a(n)$ constant as in phase modulation,

$$\begin{aligned} U_{01}(x) = {} & \{\sum \cos [\omega_c t - \psi_c(n) - \varphi_c(x - n) + \theta]\bar{P}_c(x - n)\} \\ & \cdot\{\sum \cos [\omega_c t - \psi_c(n + 1) - \varphi_c(x - n)]\bar{P}_c(x - n)\}, \end{aligned} \tag{31}$$

where as before the summations are between $n = -\infty$ and $n = \infty$.

After elimination of high-frequency components present in (31) by low-pass filtering and omitting a factor of one-half, the resultant baseband output can be written

$$U_{01}(x) = \sum S_n(x)\bar{P}_c(x - n), \tag{32}$$

$$U_{01}(0) = \sum S_n(0)\bar{P}_c(-n), \tag{33}$$

in which the summations are between $n = -\infty$ and ∞ and

$$\begin{aligned} S_n(x) = {} & \sum_{m=-\infty}^{\infty} \bar{P}_c(x - m) \cos [\psi_c(-n) - \psi_c(-m + 1) \\ & + \varphi_c(x - n) - \varphi_c(x - m) - \theta], \end{aligned} \tag{34}$$

$$\begin{aligned} S_n(0) = {} & \sum_{m=-\infty}^{\infty} \bar{P}_c(-m) \cos [\psi_c(-n) - \psi_c(-m + 1) \\ & + \varphi_c(-n) - \varphi_c(-m) - \theta]. \end{aligned} \tag{35}$$

The desired output is represented by the term in (33) for $n = 0$ and is

$$U_{01}^{0}(0) = S_0(0)\bar{P}_c(0) \tag{36}$$

where, in accordance with (35),

$$\begin{aligned} S_0(0) = {} & \sum_{m=-\infty}^{\infty} \bar{P}_c(-m) \cos [\psi_c(0) - \psi_c(-m + 1) \\ & + \varphi_c(0) - \varphi_c(-m) - \theta]. \end{aligned} \tag{37}$$

In the absence of intersymbol interference $\bar{P}_c(-m) = 0$ for $m \neq 0$ and

$$S_0^{0}(0) = \bar{P}_c(0) \cos [\psi_c(0) - \psi_c(1) - \theta] \tag{38}$$

so that (36) become

$$U_{01}^{0}(0) = \bar{P}_c^{2}(0) \cos [\psi_c(0) - \psi_c(1) - \theta]. \tag{39}$$

It will be recognized that this expression is of the same form as (27) except that $\psi_c(0)$ is replaced by the phase change $\psi_c(0) - \psi_c(1)$ between two successive sampling instants. In the particular case of four-phase transmission the phase θ may be chosen as, say, $\theta = 0$ or $\pi/4$, in which case the outputs of the two product demodulators would be as indicated in Table I for phase modulation with synchronous detection, except that $\psi_c(0)$ is replaced by the phase difference $\psi_c(0) - \psi_c(1)$.

With a signal of bandwidth ω_s the high-frequency part of (31) will have a lowest frequency $2(\omega_c - \omega_s)$, since it is the product of two high-frequency components each of lowest frequency $\omega_c - \omega_s$. The baseband signal represented by (32) will have a maximum frequency $2\omega_s$, so that a flat low-pass filter of minimum bandwidth $2\omega_s$ is required to avoid distortion of the baseband signal. In order that the filter also eliminate the high-frequency components in the demodulator output, it is necessary that $2(\omega_c - \omega_s) \geqq 2\omega_s$ or $\omega_c \geqq 2\omega_s$. With synchronous detection it was necessary that $\omega_c \geqq \omega_s$.

2.7 *Binary FM with Frequency Discriminator Detection*

In frequency modulation $a(n)$ in (14) is constant and $\psi_c(n)$ varies with time so that the time derivative of $[\omega_c t - \psi_c(n)]$ represents a variable frequency. Pulse transmission without intersymbol interference over a channel of the same bandwidth as required for double-sideband AM is in this case possible for certain ideal amplitude and phase characteristics of the channels, as shown elsewhere [Ref. 1, Section 5]. The formulation is here modified to include any amplitude and phase characteristic of the channels.

It will be assumed that a space is represented by a frequency $\omega_0 - \tilde{\omega}$ and a mark by a frequency $\omega_0 + \tilde{\omega}$. Discontinuity in a transition from mark to space can then be avoided for rectangular modulating pulses of duration T provided,

$$\tilde{\omega}T = k\pi, \qquad k = 1,2,3. \tag{40}$$

In a system of minimum bandwidth $k = 1$, and in this case intersymbol interference can be avoided with a channel band no wider than required for double-sideband AM.

When a mark is preceded and followed by a space during the nth pulse interval, the envelope of the resultant carrier pulse is obtained with $t_0 = (t + nT)$ in Equation (23) of Ref. 1 and becomes

$$\begin{aligned} \bar{E}_{sm}{}^{0}(t + nT) &= 2 \cos \tilde{\omega}(t + nT) \\ &= (-1)^n 2 \cos \tilde{\omega}t, \end{aligned} \tag{41}$$

where the last relation follows from (40) with $k = 1$.

The resultant carrier pulse during interval 0 is of the general form

$$\begin{aligned} P(t) &= \cos(\omega_0 t + \varphi_0)(-1)^n R_0(t - nT) \\ &\quad + \sin(\omega_0 t + \varphi_0)(-1)^n Q_0(t - nT), \end{aligned} \tag{42}$$

where t is the time from the midpoint of interval 0.

When $\psi(-\tilde{\omega})$ is the phase distortion* at the frequency $\omega_0 - \tilde{\omega}$, Equation (34) of Ref. 1 is modified into

$$\begin{aligned} E(t) &= -\cos(\omega_0 t + \varphi_0)[A(-\tilde{\omega})\cos y - (-1)^n R_0(t - nT)] \\ &\quad + \sin(\omega_0 t + \varphi_0)[A(-\tilde{\omega})\sin y - (-1)^n Q_0(t - nT)], \end{aligned} \tag{43}$$

where

$$y = \tilde{\omega} t + \psi(-\tilde{\omega}). \tag{44}$$

When a sequence of marks and spaces is transmitted, the resultant wave at the detector input becomes

$$\begin{aligned} W_0(t) &= \cos(\omega_0 t + \varphi_0)[A(-\tilde{\omega})\cos y - \alpha_0(x)] \\ &\quad + \sin(\omega_0 t + \varphi_0)[A(-\tilde{\omega})\sin y - \beta_0(x)], \end{aligned} \tag{45}$$

where

$$\alpha_0(x) = \sum_{n=-\infty}^{\infty} (-1)^n a(n) R_0(x - n), \tag{46}$$

$$\beta_0(x) = \sum_{n=-\infty}^{\infty} (-1)^n a(n) Q_0(x - n), \tag{47}$$

in which the notation is in accordance with (13), $x = t/T$ and $y = \pi x + \psi(-\tilde{\omega})$.

The phase of the wave (45) is given by

$$\tan \Psi_0(t) = -\frac{\sin y - \mu\beta_0(x)}{\cos y - \mu\alpha_0(x)}, \tag{48}$$

where

$$\mu = 1/A(-\tilde{\omega}) \tag{49}$$

Expression (41) of Ref. 1 for a single pulse is replaced by the following for the demodulated pulse train at $x = t/T$:

$$\begin{aligned} U_0(x) &= \frac{\mu}{2D}\left[\mu(\alpha_0^2 + \beta_0^2) - \alpha_0 \cos y - \beta_0 \sin y\right. \\ &\quad \left. - \frac{1}{\tilde{\omega}}(\alpha_0' \sin y - \beta_0' \cos y) - \frac{\mu}{\tilde{\omega}}(\beta_0'\alpha_0 - \alpha_0'\beta_0)\right], \end{aligned} \tag{50}$$

* As in Ref. 1, the linear component of the phase characteristic is disregarded since it only represents a fixed transmission delay.

where

$$D = 1 + \mu^2(\alpha_0^2 + \beta_0^2) - 2\mu(\alpha_0 \cos y + \beta_0 \sin y), \qquad (51)$$

in which

$$\alpha_0 = \alpha_0(x), \quad \beta_0 = \beta_0(x), \quad \alpha_0' = d\alpha_0/dt, \quad \beta_0' = d\beta_0/dt.$$

2.8 *Signal-to-Noise Ratios in Binary FM*

Since binary FM with frequency discriminator detection is a nonlinear modulation method, determination of the optimum signal-to-noise ratio at the detector input for a given error probability presents a very difficult analytical problem, at least when consideration is given to minimum bandwidth requirements together with appropriate shaping of bandpass and postdetection low-pass filters. In Ref. 1 these various factors were taken into account, but the signal-to-noise ratios at sampling instants were evaluated on the approximate basis of a steady state carrier representing a continuing space or mark and a relatively high signal-to-noise ratio. On this basis it turned out that, in the absence of a postdetection low-pass filter, binary FM would have a disadvantage in signal-to-noise ratio of about 4.5 db compared to an optimum bipolar AM or phase reversal system. This would be reduced to about a 1.5-db disadvantage by addition of an optimum low-pass filter. The analysis further indicated that, for a specified postdetection low-pass filter, there would be an optimum division of shaping between the transmitting and receiving bandpass filters that would give a slight advantage in signal-to-noise ratios over an optimum bipolar AM system. In view of the approximations involved, the above analysis does not prove that such an advantage exists. Rather, it is probable that optimum bipolar AM has some advantage in signal-to-noise ratio over optimum binary FM. This is indicated by other analyses that do not assume a high signal-to-noise ratio but introduce other approximations in that they do not consider frequency discriminator detection or the shaping of band-pass filters or the effect of a postdetection low-pass filter.

It is well known that an approximation is involved in assuming high signal-to-noise ratios and thus ignoring the breaking phenomenon in FM. Moreover, even in the absence of intersymbol interference, it is an approximation to assume a steady state carrier over a short sampling interval, regardless of the transmitted code, as shown below.* Referring to Equation (202) of Ref. 1, random noise at the detector input can be written in the form

$$e_i(t) = r_i(t) \cos(\omega_0 t + \varphi_0) + q_i(t) \sin(\omega_0 t + \varphi_0). \qquad (52)$$

* This was shown by A. P. Stamboulis.

When this noise is combined with the signal as given by (43), equation (48) for the phase in the presence of interference becomes

$$\tan \psi_{0,i} = -\frac{\sin y - \mu[\beta_0(x) + q_i(x)]}{\cos y - \mu[a_0(x) + r_i(x)]}, \tag{53}$$

where $x = t/T$.

In the absence of intersymbol interference at sampling instants, $a_0(0) = 0$ or 1, $\beta_0(0) = 0$, $y = 0$ and $\mu = 2$. In this case appropriate modification of (50) gives for the demodulated signal plus noise at sampling instants

$$U_0(U) + U_i(0) = \frac{(\alpha_0 + r_i)[2(\alpha_0 + r_i) - 1] + q_i'/\tilde{\omega} + 2\alpha_0' q_i/\tilde{\omega}}{[2(\alpha_0 + r_i) - 1]^2}, \tag{54}$$

where $r_i = r_i(0)$, $q_i = q_i(0)$ and $q_i' = dq_i(t)/dt$ for $t = 0$.

If $r_i \ll 1$, the last equation is approximated by

$$U_0(0) + U_i(0) \cong \alpha(0) - r_i + q_i'/\tilde{\omega} + 2\alpha_0' q_i/\tilde{\omega}, \tag{55}$$

where $U_0(0) = \alpha(0) = 0$ for space and 1 for mark, and the interfering voltage after demodulation is

$$U_i(0) \cong -r_i + q_i'/\tilde{\omega} + 2\alpha_0' q_i/\tilde{\omega} \tag{56}$$

The first two terms represent the conventional approximation for a continuing mark or space and a high signal-to-noise ratio.

In order to neglect the third term in (56) it is necessary that $\alpha_0'(0) = 0$. This is not the case except for a continuing space, a continuing mark or a mark preceded and followed by a continuing space. For other combinations of transmitted pulses there is some contribution from the third term. In the particular case of a raised cosine pulse spectrum, as considered herein, the maximum effect for a random pulse train is less than 0.15 db and can thus be ignored. For narrower pulse spectra the effect may be appreciably greater.

In the analysis that follows, transmission impairments from intersymbol interference owing to phase distortion will be evaluated on the same basis for FM as for the other modulation methods, although the approximations involved may be somewhat greater.

2.9 *Slicing Levels and Noise Margins*

As indicated by the preceding derivations, the demodulated wave is related to the received carrier wave $W_0(x)$ in a manner that depends on the carrier modulation and detection method. In general the demodulated wave at sampling instants may assume a number of different

amplitudes. Let $U^{(s)}$ designate the demodulated wave for one particular amplitude or state a_s of the transmitted signal and $U^{(s+1)}$ the demodulated wave at a sampling instant for an adjacent amplitude or state s_{s+1} of the transmitted signal. There will then be a certain sequence of transmitted pulses for which a maximum value ${U_{\max}}^{(s)}$ is obtained, owing to intersymbol interference, and also a certain sequence resulting in a minimum value ${U_{\min}}^{(s+1)}$. If there is equal probability of a_s and a_{s+1} and of positive and negative noise voltages, the optimum level for distinction between $U^{(s)}$ and $U^{(s+1)}$ is

$$L_0^{(s)} = \tfrac{1}{2}[{U_{\min}}^{(s+1)} + {U_{\max}}^{(s)}]. \tag{57}$$

In the presence of $U^{(s)}$ the margin for distinction from $U^{(s+1)}$ is

$$M^{(s)} = L_0^{(s)} - U^{(s)} \tag{58}$$

and in the presence of $U^{(s+1)}$ the margin for distinction from $U^{(s)}$ is

$$M^{(s+1)} = U^{(s+1)} - L_0^{(s)}. \tag{59}$$

The minimum margins are obtained with $U^{(s)} = {U_{\max}}^{(s)}$ and with $U^{(s+1)} = {U_{\min}}^{(s+1)}$ in (58) and (59). The minimum margins thus become

$${M_{\min}}^{(s+1)} = {M_{\min}}^{(s)} = \tfrac{1}{2}[{U_{\min}}^{(s+1)} - {U_{\max}}^{(s)}]. \tag{60}$$

For sequences of marks and spaces, or other signal patterns, such that the minimum margins for distinction between adjacent signal states are obtained, an error will occur if the noise voltage at the sampling instant exceeds $M_{\min}$ in amplitude and has the appropriate polarity. (Polarity is immaterial except for the two extreme signal states.) For other signal patterns the tolerable amplitude of the noise voltage is greater. The value of $M_{\min}$ relative to the value in the absence of intersymbol interference thus gives the maximum transmission impairment. The average impairment obtained by considering various pulse train patterns and the corresponding values of $M^{(s)}$ and $M^{(s+1)}$ as given by (58) and (59) will be less, as discussed below.

2.10 *Evaluation of Transmission Impairments*

By way of illustration it will be assumed that all values of M between $M_{\min}$ and a maximum value $M_{\max}$ are equally probable, and that the noise has a gaussian amplitude distribution. With a given fixed value of M the probability of an error can be written as

$$p_e = \tfrac{1}{2} \operatorname{erfc} (aM) \tag{61}$$

where erfc = 1 − erf is the error function complement and a is a factor that depends on the ratio of signal power to noise power.

Considering all noise margins between the limits mentioned above, the average error probability becomes

$$\bar{p}_e = \frac{1}{M_{\max} - M_{\min}} \frac{1}{2} \int_{M_{\min}}^{M_{\max}} \operatorname{erfc}(aM)\,dM$$

$$= \frac{1}{2} \frac{1}{M_{\max} - M_{\min}} \left[M_{\max} \operatorname{erfc} B - M_{\min} \operatorname{erfc} A + \frac{1}{a\sqrt{\pi}} (e^{-A^2} - e^{-B^2}) \right], \tag{63}$$

where

$$A = a\,M_{\min},$$

$$B = a\,M_{\max}.$$

With

$$M_{\max} = k\,M_{\min}, \tag{64}$$

(63) becomes

$$\bar{p}_e = \frac{1}{2} \frac{1}{k-1} \left[k \operatorname{erfc}(kA) - \operatorname{erfc} A + \frac{1}{A\sqrt{\pi}} (e^{-A^2} - e^{-k^2A^2}) \right]. \tag{65}$$

For $k = 1$, the latter expression conforms with (61).

The maximum error probability would be obtained by considering a fixed noise margin equal to $M_{\min}$ and would be

$$\hat{p}_e = \tfrac{1}{2} \operatorname{erfc} A. \tag{66}$$

The error committed in assuming $M_{\min}$ can be determined by writing $\bar{p}_e$ as given by (65) in the form

$$\bar{p}_e = \tfrac{1}{2} \operatorname{erfc}(cA), \tag{67}$$

where $c \geqq 1$ is so chosen that (67) equals (65).

The average noise margin is then

$$\bar{M} = c\,M_{\min} \tag{68}$$

By way of numerical illustration let A be so chosen that $\hat{p}_e$ as given by (66) in one case is 10^{-4} and in another case 10^{-5}. The results given in Table II are then obtained from (65) and (67).

As will be shown in a later section, a factor $k = 3$ may correspond to

TABLE II — RATIO $c = \bar{M}/M_{min}$ FOR EQUAL PROBABILITY OF ALL NOISE MARGINS BETWEEN M_{min} AND $M_{max} = k\,M_{min}$ FOR NOISE WITH A GAUSSIAN AMPLITUDE DISTRIBUTION

	$A = 2.62$; $\hat{p}_e = 10^{-4}$			$A = 3.0$; $\hat{p}_e = 10^{-5}$		
$k =$	1	2	3	1	2	3
p_e	10^{-4}	10^{-5}	5×10^{-6}	10^{-5}	1.5×10^{-6}	7.5×10^{-7}
c	1	1.15	1.17	1	1.1	1.12
c (in db)	0	1.2	1.4	0	0.8	1.0

a transmission impairment of about 10 db based on the minimum noise margin, whereas the actual impairment would be 1.4 db less for an error probability of 10^{-4}, about 1 db less for an error probability 10^{-5}. For an error probability of 10^{-6} or less the error committed in evaluating transmission impairments on the basis of the minimum noise margin can be disregarded. This also applies for greater error probabilities when the transmission impairment based on the minimum noise margin is small, in which case $k < 2$.

III. SYNCHRONOUS AM AND PM

3.1 *General*

Amplitude modulation can be used in conjunction with envelope detection and synchronous detection. The former method is simplest from the standpoint of implementation, but synchronous detection, also referred to as homodyne and coherent detection, affords an improvement in signal-to-noise ratio. Since synchronous detection is also the simplest method from the standpoint of analysis, it will be considered here, except for a comparison of envelope and synchronous detection for binary double-sideband AM.

Amplitude modulation in general implies several pulse amplitudes, and can be used with double-sideband and with vestigial-sideband transmission. The particular case of bipolar binary AM with synchronous detection is equivalent to two-phase modulation.

With amplitude modulation and synchronous detection it is possible to transmit pulse trains on two carriers at quadrature with each other, and under certain idealized conditions to avoid mutual interference. The special case of bipolar binary AM on each of the two carriers is equivalent to four-phase modulation.

The signal-to-noise ratio as related to error probability is discussed elsewhere (Ref. 1, Section XVIII) for various optimized binary AM or

PM systems on the premise of ideal synchronous detection. Ideal synchronous detection for AM or PM as assumed here can in principle be approached without penalty in signal-to-noise ratio, by various methods of implementation. For example, a demodulating wave for a product demodulator can be derived with the aid of a resonator of sufficiently narrow bandwidth (high Q) tuned to the carrier frequency, or the second or the fourth harmonic thereof, depending on the particular method and on whether two-phase or four-phase modulation is used. A demodulating wave can also be supplied from an oscillator at the receiving end, the phase of which would be controlled by comparison with that of the carrier of the received signal. Such phase-locked oscillator methods have been devised for analog signal transmission by suppressed carrier double-sideband AM[5] and vestigial-sideband AM.[6] With any one of the above methods, noise in the demodulating wave would be virtually absent, as would the effect of phase distortion in the channel. Actually some penalty in signal-to-noise ratio as compared to ideal synchronous detection would be incurred, owing to unavoidable fluctuations in the amplitude and phase of the demodulating wave, resulting from the finite bandwidth of the resonators and mistuning, or from imperfect oscillator control. A common property of these methods is that a rather long time, as measured in pulse intervals, is required to establish a demodulating wave, if the above fluctuations in amplitude and phase are to be held within tolerable limits. This may be a disadvantage in certain applications, which in the case of phase modulation can be overcome by differential phase detection, in exchange for a penalty in signal-to-noise ratio resulting from the presence of both noise and phase distortion in the demodulating wave, as discussed in Section IV.

A general formulation is given here of intersymbol interference and resultant maximum transmission impairment as related to the carrier pulse transmission characteristic, together with illustrative applications to the particular cases of linear and quadratic delay distortion. The formulation is, however, applicable to any given gain and phase deviation over the channel band, provided the carrier pulse transmission characteristic has been determined, which in general would entail Fourier integral evaluation with the aid of computers.

3.2 *Synchronous AM and Two-Phase Modulation*

With synchronous detection (22) applies, or alternately, with $x = 0$ and $U_0(0) = U$,

$$U = a(0)R_c(0) + \sum_{n=1}^{\infty} [a(-n)R_c(n) + a(n)R_c(-n)]. \qquad (69)$$

The following notation will be used:

$$r_c^{+} = \sum_{n=1}^{\infty} [R_c^{+}(n) + R_c^{+}(-n)], \tag{70}$$

$$r_c^{-} = \sum_{n=1}^{\infty} [R_c^{-}(n) + R_c^{-}(-n)], \tag{71}$$

where R_c^{+} designates positive values of R_c and R_c^{-} absolute values when R_c is negative.

Let there be l different amplitude levels, between a minimum amplitude $a_{\min}$ and a maximum amplitude $a_{\max}$. When a pulse of amplitude $a_s = a_s(0)$ is transmitted, the maximum value of (69) is

$$U_{\max}^{(s)} = a_s R_c(0) + a_{\max} r_c^{+} - a_{\min} r_c^{-}. \tag{72}$$

For the next higher pulse amplitude $a_{s+1} = a_s + (a_{\max} - a_{\min})/(l-1)$, the minimum value of (69) is

$$U_{\min}^{(s+1)} = a_{s+1} R_c(0) + a_{\min} r_c^{+} - a_{\max} r_c^{-}. \tag{73}$$

The minimum noise margin is, in accordance with (60),

$$M_{\min} = \frac{a_{\max} - a_{\min}}{2} \left[\frac{R_c(0)}{l-1} - r_c^{+} - r_c^{-} \right]. \tag{74}$$

In the absence of intersymbol interference $r_c^{+} = 0$, $r_c^{-} = 0$ and $R_c(0) = R_c^{(0)}(0)$, so that

$$M^0 = \frac{a_{\max} - a_{\min}}{2} \frac{R_c^0(0)}{l-1}. \tag{75}$$

The value of $M_{\min}$ as given by (74) is smaller than M^0 in the absence of intersymbol interference by the factor

$$\eta_{\min} = \frac{R_c(0)}{R_c^0(0)} \left[1 - (l-1) \frac{r_c^{+} + r_c^{-}}{R_c(0)} \right] \tag{76}$$

$$= \frac{R_c(0)}{R_c^0(0)} \left[1 - (l-1) \frac{\bar{r}_c}{R_c(0)} \right], \tag{77}$$

where

$$\bar{r}_c = \sum_{n=1}^{\infty} [\bar{R}_c(-n) + \bar{R}_c(n)], \tag{78}$$

in which $\bar{R}_c$ designates the absolute value of R_c.

The factor $R_c(0)/R_c^0(0)$ represents the transmission impairment owing to reduction in pulse amplitude at sampling instants. The summation

term represents transmission impairments owing to intersymbol interference.

Relation (77) applies regardless of the polarity of the transmitted pulses and for both symmetrical (double sideband) and asymmetrical (vestigial sideband) systems. The special case $l = 2$ and $a_{\min} = -a_{\max}$ represents binary bipolar AM, which can also be regarded as two-phase transmission.

3.3 *Quadrature Carrier AM and Four-Phase Modulation*

With synchronous detection it is possible under certain ideal conditions to transmit signals on two carriers at quadrature without mutual interference. In general, however, the quadrature component in (21) will in this case give rise to interference and (69) is replaced by

$$U = a(0)R_c(0) + \sum_{n=1}^{\infty} [a(-n)R_c(n) + a(n)R_c(-n)] + b(0)Q_c(0) + \sum_{n=1}^{\infty} [b(-n)Q_c(n) + b(n)Q_c(-n)], \tag{79}$$

where $b(n)$ are the pulse amplitudes in the quadrature system.

For equal differences between maximum and minimum amplitudes in the two systems, i.e., $a_{\max} - a_{\min} = b_{\max} - b_{\min}$, (77) is replaced by

$$\eta_{\min} = \frac{R_c(0)}{R_c^0(0)} \left\{ 1 - \frac{l-1}{R_c(0)} [Q_c(0) + \bar{r}_c + \bar{q}_c] \right\}, \tag{80}$$

where $\bar{r}_c$ is defined by (78), and similarly

$$\bar{q}_c = \sum_{n=1}^{\infty} \left[\bar{Q}_c(-n) + \bar{Q}_c(n) \right], \tag{81}$$

where $\bar{Q}_c$ designates the absolute values of Q_c.

In general the phase of the demodulating carrier can be so chosen that $Q_c(0) = 0$, as is demonstrated later.

Expression (80) applies regardless of pulse polarities in the two quadrature systems. The special case of two binary bipolar AM systems, i.e., $l = 2$ and $a_{\min} = b_{\min} = -a_{\max} = -b_{\max}$ can also be regarded as four-phase transmission.

3.4 *Even Symmetry Pulse Spectrum and Delay Distortion*

When the spectrum of a received pulse at the detector input has even symmetry about the carrier frequency, and the phase characteristic has odd symmetry (even symmetry delay distortion), the quadrature com-

ponents $Q_c(n)$ vanish (see the Appendix). In this case (77) and (80) are identical, so that there is no mutual interference between pulse trains transmitted on two carriers at quadrature. In this special case it is thus possible by quadrature carrier AM to realize a two-fold increase in pulse transmission rate, without increased intersymbol interference. An alternative means to the same end is to use vestigial sideband transmission, as discussed below.

Let T be the pulse interval in double-sideband AM, in which case the pulse interval in vestigial-sideband AM would be $T' = T/2$. Returning to (10) and (11) let $R_0(t)$ be the in-phase component in double-sideband AM, and let $Q_0(t) = 0$ for an amplitude spectrum with even symmetry about ω_0 and a phase characteristic with odd symmetry. Let ω_y be the carrier frequency from midband in vestigial-sideband transmission. By appropriate choice of ω_y it is possible to make $\omega_y T' = \pi/2$, in which case $\cos \omega_y T' = 0$, $\sin \omega_y T' = 1$. The following relations are thus obtained:

At even sampling points, i.e., $m = 0,2,4,6, \cdots$,

$$R_{c,0}(mT') = (-1)^{m/2} R_0(mT') = (-1)^{m/2} R_0(mT/2),$$
$$Q_{c,0}(mT') = 0. \tag{82}$$

At odd sampling points, i.e., $m = 1,3,5,7, \cdots$,

$$R_{c,0}(mT') = 0,$$
$$Q_{c,0}(mT') = (-1)^{(m-1)/2} R_0(mT') = (-1)^{(m-1)/2} R_0(mT/2). \tag{83}$$

In accordance with the above relations, at even sampling points only the in-phase components are present and are the same as in double-sideband AM. At odd sampling points the quadrature components are present, but are eliminated with synchronous detection and need not be considered.

In summary, when the amplitude spectrum at the detector input has even symmetry about the midband frequency, and the phase characteristic has odd symmetry, relation (77) applies for double-sideband AM, quadrature double-sideband AM, vestigial-sideband AM, as well as special cases thereof, such as two-phase and four-phase modulation.

In the next section numerical results are given for the special case of a raised cosine spectrum at the detector input with quadratic delay distortion about the midband frequency.

3.5 *Raised Cosine Spectrum and Quadratic Delay Distortion*

In the following numerical illustration, the spectrum at the detector input will be assumed to have a raised cosine shape, as shown in Fig. 1.

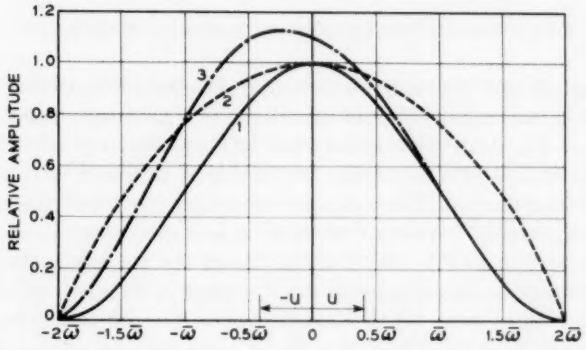


Fig. 1 — Raised cosine pulse spectrum and transmission-frequency characteristic of channel for double- and vestigial-sideband AM. Curve 1: Spectrum of carrier pulse envelope at detector input (channel output); $S(u)/S(0) = \cos^2(\pi u/4\bar{\omega})$. Curve 2: Transmission-frequency characteristic of channel with rectangular transmitted carrier pulses of duration $T = \pi/\bar{\omega}$ and carrier at midband; $A(u)/A(0) = (\pi u/4\bar{\omega})/\tan(\pi u/4\bar{\omega})$. Curve 3: Transmission frequency characteristic of channel with rectangular transmitted carrier pulses of duration $T/2$ and carrier at $u = \bar{\omega}$; $A(u)A(0) = \cos^2(\pi u/4\bar{\omega})\{[(u - \bar{\omega})/4\bar{\omega}]/\sin[\pi(u - \bar{\omega})/4\bar{\omega}]\}$.

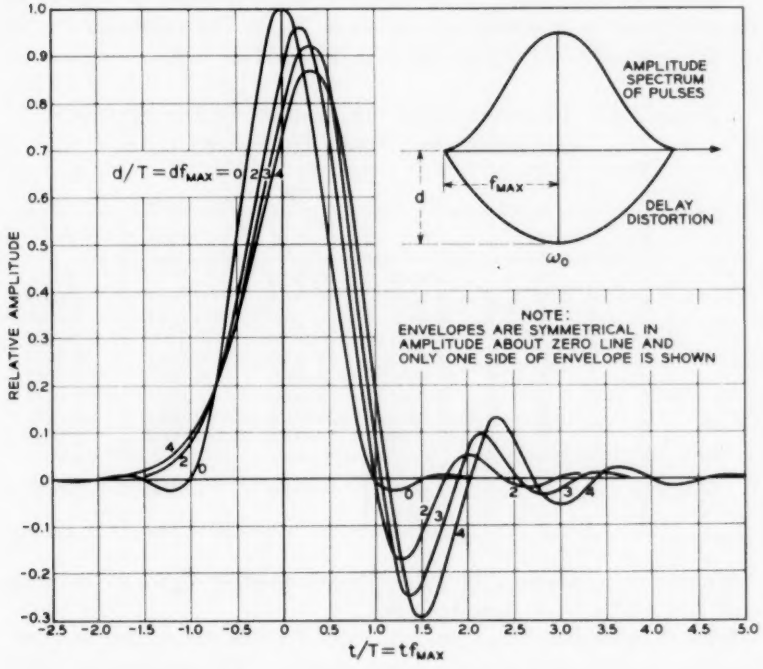


Fig. 2 — Carrier pulse transmission characteristic for raised cosine spectrum as in Fig. 1 and quadratic delay distortion.

The shape of the transmission-frequency characteristic of the channel required to this end depends on the shape of the transmitted pulses. It is shown in Fig. 1 for rectangular modulating pulses, with the carrier at midband and also with the carrier to one side of midband, as in vestigial-sideband transmission. These characteristics, together with the optimum division of shaping between transmitting and receiving filters, are discussed in Section XIV of Ref. 1. Even though the amplitude characteristics of the detector input spectra are the same in double- and vestigial-sideband transmission, it is necessary to use different shaping of transmitting filters, as indicated in Fig. 1, since the rectangular modulating pulses have different carrier frequencies and different durations.

The phase characteristic is assumed to contain a linear component, together with phase distortion component varying as the third power of frequency from midband, which corresponds to delay distortion increasing as the second power of frequency from midband, as indicated in Fig. 2. The function $R_0(t/T) = R_0(x + n)$ for this case has been determined by numerical integration, as discussed in the Appendix. It is given in Table XIX of the Appendix and shown in Fig. 2. The values for $x = 0$, i.e., integral values of t/T are given in Table III.

TABLE III — FUNCTION $R_0(n)$ FOR RAISED COSINE SPECTRUM AND QUADRATIC DELAY DISTORTION

$n = t/T$	d/T				
	0	1	2	3	4
−3	0	−0.0006	0.0025	0	0
−2	0	−0.0013	0.0011	0.0017	0.0028
−1	0	0.0467	0.0756	0.0891	0.0986
0	1	0.9633	0.8795	0.7956	0.7336
1	0	−0.0341	−0.0098	0.0827	0.2045
2	0	0.0196	0.0543	0.0655	0.0142
3	0	0.0044	0.0020	−0.0231	−0.0584
4	0	0.0014	−0.0014	−0.0087	−0.0037
5	0	0.0006	−0.0008	−0.0022	0.0040
$\sum_{-\infty}^{\infty} R_0(n)$	1.0	1.0000	1.0004	1.0006	0.9957

With exact evaluation of $R_0(n)$ the summation $\sum R_0(n)$ between $n = -\infty$ and $n = \infty$ should equal 1.

With the values of $R_0(n)$ given in Table III, the values of $\bar{r}_c = \bar{r}_0$ obtained from (78), and of η_{min} as obtained from (77) are given in Table IV.

These factors are shown in Fig. 3 and apply for double- and vestigial-

TABLE IV — FACTOR η_{min} FOR RAISED COSINE SPECTRUM AND QUADRATIC DELAY DISTORTION FOR SYNCHRONOUS AM WITH l AMPLITUDE LEVELS*

d/T	0	1	2	3	4
r_0	0	0.109	0.148	0.273	0.385
$l = 2$	1	0.855	0.732	0.523	0.347
$l = 3$	1	0.746	0.585	0.250	−0.037
$l = 4$	1	0.637	0.437	−0.053	—
$l = 5$	1	0.529	0.300	−0.296	—

* Results also apply with envelope detection (Section 3.9)

sideband AM and for quadrature double-sideband AM, and special cases thereof, such as two-phase and four-phase transmission. Since the quadrature component is absent the factors also apply for double-sideband AM with envelope detection. It should be noted that T in all cases is the pulse interval in double-sideband AM, which is twice the pulse interval

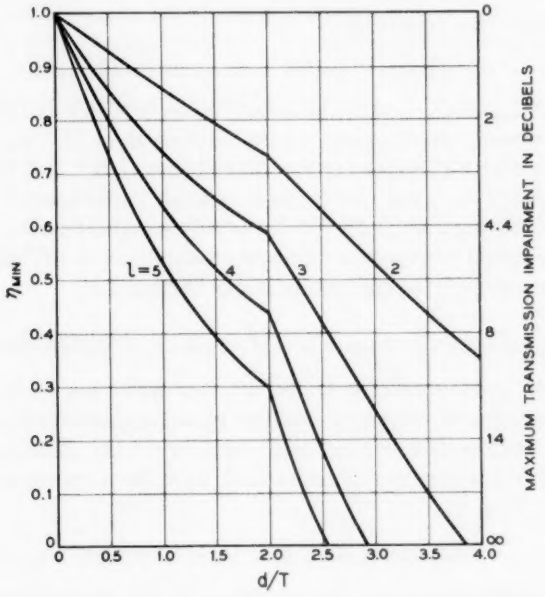


Fig. 3 — Factor η_{min} for raised cosine pulse spectrum and quadratic delay distortion as in Fig. 2, for AM systems employing synchronous detection and l pulse amplitudes. Factor applies for double- and vestigial-sideband AM and quadrature double-sideband AM.

in vestigial-sideband AM or twice the combined pulse interval in quadrature double-sideband AM.

In the above evaluation it was assumed that the pulses were sampled at $t = 0$, which is not at the peak of a pulse except for $d/T = 0$. For $d/T = 4$ the pulse peak is nearly at $t/T = x = 0.25$. Sampling at $x = 0.25$ gives, for $l = 2$, $\eta_{\min} = 0.356$ rather than 0.347.

The factor $\eta_{\min}$ expressed in decibels, as in Fig. 3, indicates the maximum transmission impairment, i.e., the maximum increase in signal-to-noise ratio required at the detector input to compensate for the effect of delay distortion. This maximum impairment would be closely approached for signal-to-noise ratios such that the error probability is sufficiently small, say less than 10^{-7}. However, for error probabilities in the range ordinarily considered the transmission impairment will be less, in accordance with the discussion in Section 2.10. For example, for $d/T = 4$, $R_0(0) = 0.734$ and $\bar{r}_0(0) = 0.385$. The maximum noise margin for $l = 2$ is in this case

$$M_{\max} = R_0(0) + \bar{r}_0(0) \cong 1.12$$

and the minimum margin is

$$M_{\min} = R_0(0) - \bar{r}_0(0) \cong 0.397.$$

For $k = M_{\max}/M_{\min} = 3.2$, the results given in Table II indicate that the transmission impairments would be less than the maximum by about 1 db and 1.4 db for error probabilities of 10^{-5} and 10^{-4} respectively. For $d/T = 4$ and $l = 2$ the maximum impairment indicated in Fig. 3 is $-20 \log_{10} M_{\min} \cong 9.2$ db, whereas impairments of about 8.2 and 7.8 db would be expected for error probabilities of 10^{-5} and 10^{-4}, on the premises underlying the evaluation in Section 2.10.

3.6 *Even Symmetry Spectrum and Odd Symmetry Delay Distortion*

When the pulse spectrum at the detector input has even symmetry about the midband frequency and the phase characteristic has a component of even symmetry, i.e., odd symmetry delay distortion, the in-phase and quadrature components both have even symmetry with respect to t. That is,

$$R_0(-t) = R_0(t); \qquad Q_0(-t) = Q_0(t) \tag{84}$$

With synchronous detection the phase of the demodulating carrier would preferably be so chosen that $Q_0(t)$ would vanish at $t = 0$, since this would give the maximum amplitude of the demodulated pulse at a sampling constant, equal to $[R_0^2(0) + Q_0^2(0)]^{\frac{1}{2}}$. For purposes of analysis

it is therefore convenient to modify the phase angle so that the quadrature component vanishes at $t = 0$. The modified quantities are related to R_0 and Q_0 by (see Appendix)

$$R_{00}(t) = \frac{R_0(0)R_0(t) + Q_0(0)Q_0(t)}{[R_0^2(0) + Q_0^2(0)]^{\frac{1}{2}}}, \tag{85}$$

$$Q_{00}(t) = \frac{R_0(0)Q_0(t) - Q_0(0)R_0(t)}{[R_0^2(0) + Q_0^2(0)]^{\frac{1}{2}}}. \tag{86}$$

In the case of double-sideband transmission (77) applies, with

$$\bar{r}_c = \bar{r}_{00} = 2\sum_{n=1}^{\infty} \bar{R}_{00}(n). \tag{87}$$

With quadrature double-sideband transmission (80) applies, with $Q_c(0) = Q_{00}(0) = 0$ and

$$\bar{q}_c = \bar{q}_{00} = 2\sum_{n=1}^{\infty} \bar{Q}_{00}(n), \tag{88}$$

where $\bar{R}_{00}$ and $\bar{Q}_{00}$ designate absolute values.

In the case of vestigial-sideband transmission at pulse intervals $T' = T/2$, the in-phase component referred to a carrier at frequency $\omega_0 + \tilde{\omega}$ is obtained from (10) and becomes

$$R_{c,00} = \cos\left(\frac{\tilde{\omega}T}{2} m\right) R_{00}(m) - \sin\left(\frac{\tilde{\omega}T}{2} m\right) Q_{00}(m). \tag{89}$$

With $\tilde{\omega}T = \pi$, i.e., $\tilde{\omega}T' = \pi/2$, (89) gives at even sampling points, $m = 0,2,4,6, \cdots$,

$$R_{c,00}{}^{(m)} = (-1)^m R_{00}(m). \tag{90}$$

At odd sampling points, $m = 1,3,5,7, \cdots$,

$$R_{c,00}{}^{(m)} = (-1)^{(m+1)/2} Q_{00}(m). \tag{91}$$

In this case (77) applies, with $R_c(0) = R_{00}(0)$ and

$$\bar{r}_c = 2\sum_{m=2,4,6\cdots}^{\infty} \bar{R}_{00}(m) + 2\sum_{m=1,3,5\cdots}^{\infty} \bar{Q}_{00}(m). \tag{92}$$

3.7 *Raised Cosine Spectrum and Linear Delay Distortion*

For the special case of a raised cosine spectrum and linear delay distortion the functions R_0 and Q_0 have been determined by numerical integration, as discussed further in the Appendix. They are given in Table XX for certain ratios d/T, where d is the difference in delay be-

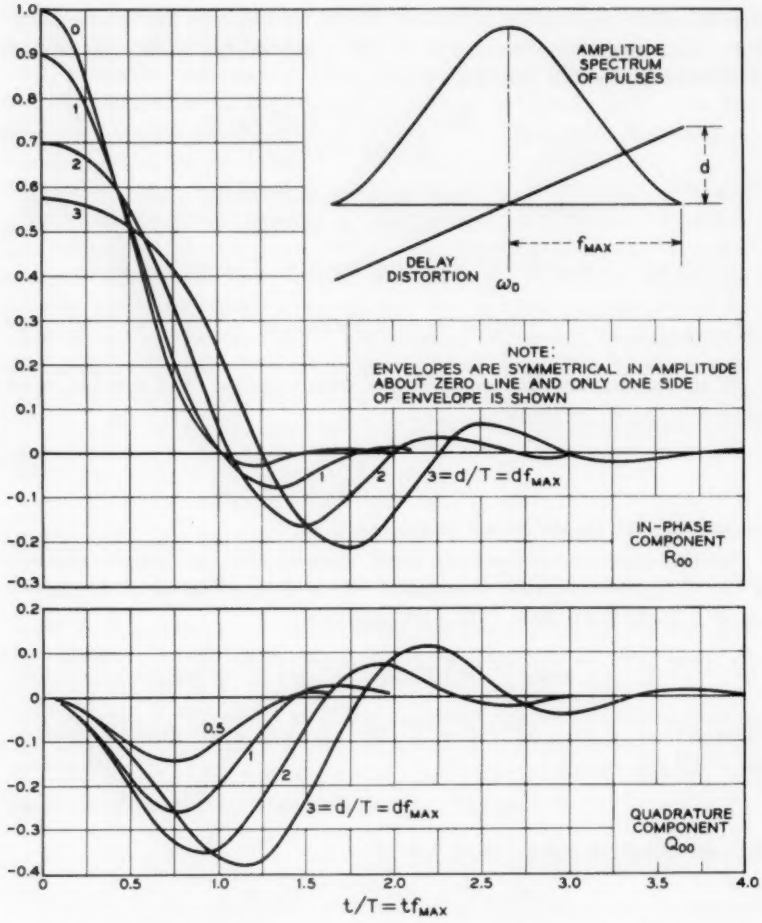


Fig. 4 — Carrier pulse transmission characteristics for raised cosine spectrum as in Fig. 1 and linear delay distortion.

tween the midband frequency and maximum sideband frequency as illustrated in Fig. 4. The modified functions R_{00} and Q_{00} are given in Table XXI and are shown in Fig. 4. For negative values of t/T, R_{00} and Q_{00} are the same as shown in Fig. 4 for positive values.

For double-sideband transmission the factors $\eta_{\min}$ given in Table V are obtained from (77), with $\bar{r}_c$ taken in accordance with (87). These factors are shown in Fig. 5. The case $l = 2$ corresponds to two-phase transmission.

TABLE V — FACTOR η_{min} FOR DOUBLE-SIDEBAND AM AND $l = 2, 3, 4$ AND 5 PULSE AMPLITUDES

l	d/T				
	0	0.5	1	2	3
2	1	0.959	0.860	0.517	−0.144
3	1	0.947	0.826	0.336	−0.860
4	1	0.935	0.792	0.155	−1.57
5	1	0.923	0.758	−0.026	—

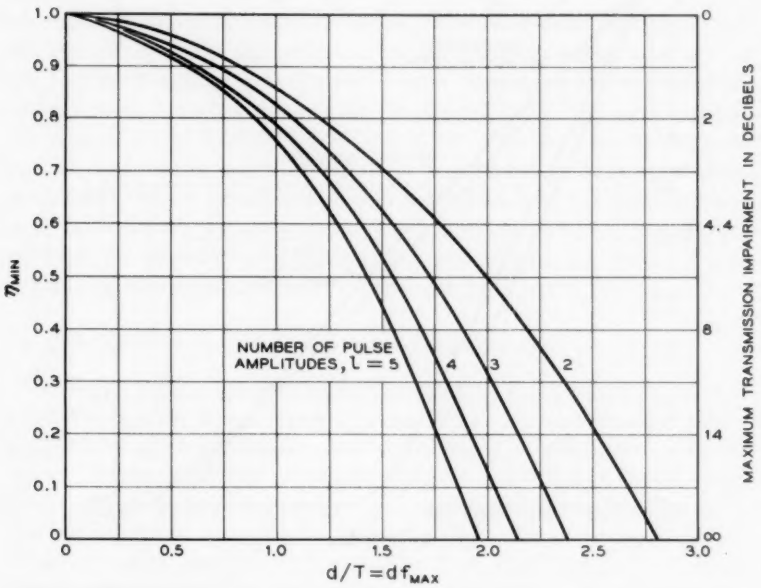


Fig. 5 — Factor η_{min} for raised cosine pulse spectrum and linear delay distortion as in Fig. 4 for double-sideband AM systems employing synchronous detection and l pulse amplitudes.

For quadrature double-sideband AM the factors in Table VI are obtained from (80) with $\bar{r}_c$ and $\bar{q}_c$ taken in accordance with (87) and (88). These factors are shown in Fig. 6. The case $l = 2$ corresponds to four-phase transmission.

For vestigial-sideband transmission the factor η_{min} is determined from (77), with $\bar{r}_c$ taken in accordance with (92). The factors given in Table VII are thus obtained.

TABLE VI — FACTOR η_{min} FOR QUADRATURE DOUBLE-SIDEBAND AM FOR $l = 2$ AND 3 PULSE AMPLITUDES

l	d/T			
	0	0.5	1	2
2	1	0.735	0.458	−0.32
3	1	0.499	0.022	−1.34

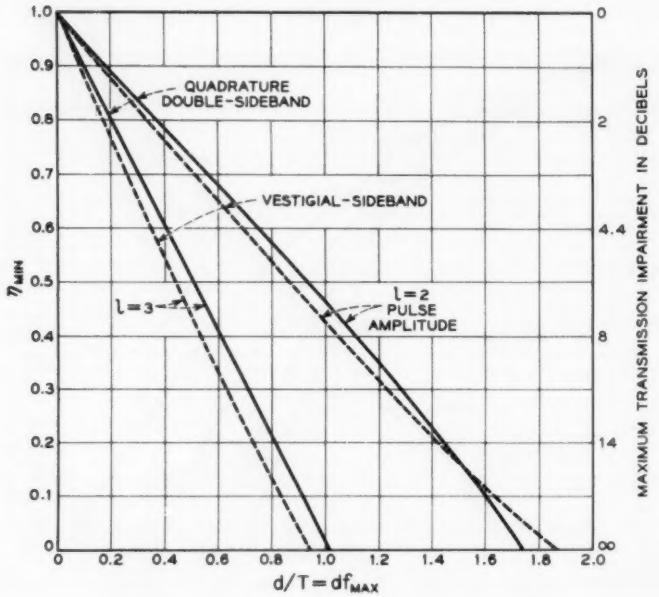


Fig. 6 — Factor η_{min} for raised cosine pulse spectrum and linear delay distortion as in Fig. 4 for quadrature double-sideband AM systems (solid lines) and vestigial-sideband AM systems (dashed lines) employing synchronous detection and $l = 2$ and 3 pulse amplitudes.

TABLE VII — FACTOR η_{min} FOR VESTIGIAL-SIDEBAND AM FOR $l = 2$ AND 3 PULSE AMPLITUDES

l	d/T			
	0	0.5	1	2
2	1	0.703	0.42	−0.061
3	1	0.435	−0.05	−0.82

3.8 *Vestigial-Sideband vs. Quadrature Double-Sideband AM*

The factors $\eta_{\min}$ for vestigial sideband AM are compared in Fig. 6 with the corresponding factors for quadrature double-sideband AM. With ideal transmission-frequency characteristics and ideal synchronous detection the two methods are equivalent from the standpoint of channel bandwidth requirements and optimum signal-to-noise ratio for a given error probability. As shown in Section 3.4, this also applies for pulse spectra at the detector input with even symmetry about the midband frequency, in the presence of delay distortion with even symmetry. The equations in Section 3.6 and the curves in Fig. 6 show that the above two methods are not equivalent in the presence of delay distortion with odd symmetry about the midband frequency. With linear delay distortion the factor $\eta_{\min}$ is, however, very nearly the same with both methods. For practical purposes quadrature double-sideband AM and vestigial-sideband AM can be regarded as equivalent with any type of delay distortion that would be expected in actual facilities. This equivalence would apply on the premise of ideal synchronous detection but not necessarily with actual implementation of synchronous detection, for the reason that the penalty in signal-to-noise ratio incurred in deriving a demodulating wave may not be the same with both methods.

3.9 *Envelope Detection vs. Synchronous Detection*

In the preceding analysis ideal synchronous detection was assumed, which permits the use of bipolar pulses. An alternative method that is simpler in implementation is envelope detection, which, however, entails the use of unipolar pulse transmission and for this reason has a certain disadvantage in signal-to-noise ratio as compared to synchronous detection.[1] In addition, transmission impairments by phase distortion may in certain cases be greater with envelope than with synchronous detection, as shown below.

When both the pulse spectrum and delay distortion have even symmetry about the carrier frequency, so that the quadrature component is absent, the effect of delay distortion is the same as with synchronous detection. The results given in Table IV thus apply also for double-sideband AM with envelope detection.

When a quadrature component is present in the carrier pulse transmission characteristic the resultant demodulated wave is in accordance with (23) given by

$$U(0) = (r_0^2 + q_0^2)^{\frac{1}{2}}, \tag{93}$$

where

$$r_0 = r_0(x) = \sum_{n=-\infty}^{\infty} a(n)R_0(x - n), \tag{94}$$

$$q_0 = q_0(x) = \sum_{n=-\infty}^{\infty} a(n)Q_0(x - n). \tag{95}$$

The modified values R_{00} and Q_{00} can be used in place of R_0 and Q_0.

Owing to the presence of both in-phase and quadrature components, it does not appear feasible to derive a simple general expression for $U_{\max}^{(s)}$ and $U_{\min}^{(s+1)}$ similar to (72) and (73). These values can, however, be determined by examining several combinations of transmitted pulses, as illustrated below for binary pulse transmission, a raised cosine pulse spectrum at the detector input and linear delay distortion. Using values of R_{00} and Q_{00} given in Table XXI of the Appendix, the results are as shown in Table VIII. Since both $R_0(t)$ and $Q_0(t)$ in this case have even symmetry about $t = 0$ the maximum effect of delay distortion is encountered for pulse trains with even symmetry about the sampling point, i.e., $a(-n) = a(n)$. Hence, only pulse trains with this property need to be considered.

From Table VIII can be obtained $W_{\max}^{(0)}$ and $W_{\min}^{(1)}$, as indicated by asterisks, together with the optimum slicing level given by (57) and the factor $\eta_{\min} = W_{\min}^{(1)} - W_{\max}^{(0)}$. These are given in Table IX.

TABLE VIII — VALUES OF $U(0)$ FOR RAISED COSINE SPECTRUM AND LINEAR DELAY DISTORTION FOR VARIOUS COMBINATIONS OF MARKS = 0 AND SPACES = 1

d/T	$a(0)$	$a(-1)$ $a(1)$	$a(-2)$ $a(2)$	$a(-3)$ $a(3)$	r_0	q_0	$U(0)$
0.5	0	1	0	0	0.036	−0.191	0.195
	0	1	1	0	0.046	−0.194	0.199*
	0	1	1	1	0.046	−0.194	0.199
	1	0	0	0	0.952	0.194	0.970*
	1	1	0	0	0.988	0.003	0.988
	1	1	1	0	0.998	0	0.998
1	0	1	0	0	0.1452	−0.3584	0.384*
	0	1	1	0	0.1676	−0.3370	0.374
	0	1	1	1	0.1688	−0.3322	0.374
	1	0	0	0	0.8309	0.3306	0.90*
	1	1	0	0	0.9761	−0.0278	0.98
	1	1	1	0	0.9985	−0.0064	0.99
2	0	1	0	0	0.5192	−0.4878	0.72*
	0	1	1	0	0.5156	0.3724	0.64
	1	0	0	0	0.5786	0.3895	0.70
	1	0	1	0	0.5020	0.5040	0.71
	1	0	0	1	0.5598	0.1380	0.68*

TABLE IX — FACTOR η_{min} WITH BINARY AM AND ENVELOPE DETECTION FOR RAISED COSINE SPECTRUM AND LINEAR DELAY DISTORTION

d/T	0	0.5	1	2
$W_{max}^{(0)}$	0	0.199	0.384	0.72
$W_{min}^{(1)}$	1	0.970	0.900	0.68
L	0.50	0.58	0.63	0.70
η_{min}	1	0.77 (0.959)†	0.60 (0.86)†	−0.04 (0.517)†

† From Table V for binary AM with synchronous detection.

It will be recognized that synchronous detection has a significant advantage over envelope detection as regards transmission impairments caused by pronounced linear delay distortion, for the reason that the effect of the quadrature component is eliminated. In general, delay distortion will have a component of even symmetry and a component of odd symmetry about the carrier frequency, in which case the quadrature component will be smaller. The advantage of synchronous detection as regards transmission impairments caused by delay distortion will then be less than indicated in Table IX. The principal advantage of synchronous detection is that it permits the use of bipolar transmission, which in the case of binary systems as considered above affords about 3 db improvement in the ratio of average signal power to average noise power for a given error probability (Ref. 1, Table VII).

In the case of vestigial-sideband transmission a pronounced quadrature component is present even in the absence of phase distortion. The advantage of synchronous detection over envelope detection is in this case significantly greater than for double-sideband transmission considered above, for the reasons that bipolar transmission can be used and quadrature component is eliminated. In the absence of phase distortion and with a raised cosine pulse spectrum at the detection input, synchronous detection has about a 9 db advantage over envelope detector in the ratio of average signal power to average noise power for a given error probability (6 db owing to elimination of quadrature component and 3 db owing to bipolar transmission).

Evaluation of transmission impairments from phase distortion is more complicated for envelope than for synchronous detection. These impairments have been determined experimentally for a binary vestigial-sideband system with an approximately raised cosine spectrum at the detector input, for linear and quadratic delay distortion and combinations thereof.[4] They are significantly greater than determined herein for synchronous detection. Hence envelope detection entails more phase equali-

zation than synchronous detection, unless a greater disparity in signal-to-noise ratio is accepted than the 9 db applying in the absence of phase distortion.

IV. PM WITH DIFFERENTIAL PHASE DETECTION

4.1 *General*

In phase modulation with differential phase detection, the demodulator output would under ideal conditions depend on changes in carrier phase between two successive pulse intervals of duration T. In its simplest and ideal form, the signal with two-phase modulation would be applied to one pair of terminals of a product demodulator, while the signal delayed by a pulse interval T would be applied to the other pair. With four-phase modulation two product demodulators are required, each with a delay network at one pair of terminals. In addition, a phase shift of 90° must be provided between all frequencies of the demodulating waves of the two demodulators, as indicated in Fig. 7. Such a phase shift over a frequency band can be realized in principle and closely approached with actual networks.[7] The modulator outputs would be applied to low-pass filters of appropriate bandwidth for elimination of high-frequency demodulation products, and the output of these would be sampled at interval T. The phase of the carrier would be indicated by the combined output as discussed in Sections 2.5 and 2.6.

With the above method it is possible with ideal channel characteristics to avoid intersymbol interference at sampling instants, without the need for a wider channel band than required with synchronous detection. However, the two methods are not in all respects equivalent from the

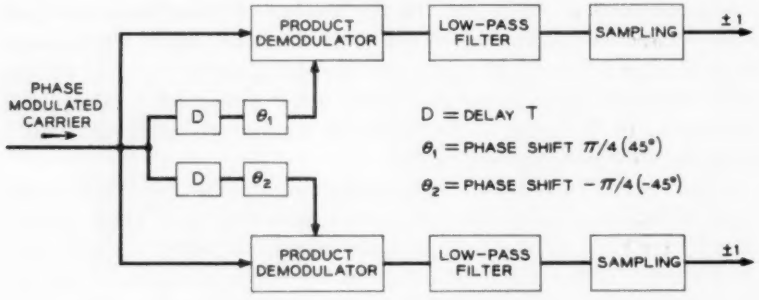


Fig. 7 — Basic demodulator arrangement for four-phase modulation with differential phase detection.

standpoint of bandwidth utilization. As discussed in Sections 2.5 and 2.6, with synchronous detection the carrier frequency must be at least equal to the maximum baseband signal frequency, whereas with differential phase detection it must exceed twice the maximum baseband signal frequency. This requirement does not impose a limitation on bandwidth utilization with differential phase modulation provided the midband frequency of the available channel is at least twice the lowest frequency, or that this condition is realized through frequency translation prior to demodulation.

With differential phase detection the demodulating wave is established without the need for a long delay (measured in pulse intervals) as required with certain other methods mentioned in Section 3.1. Moreover, a substantial fluctuation in carrier phase can be tolerated, since only the difference in phase between adjacent pulses need be considered. These advantages are realized in exchange for a penalty in signal-to-noise ratio as compared to ideal synchronous detection owing to the presence of noise in the demodulating wave. For very small error probabilities, and assuming ideal implementation in all respects, this impairment is about 1 db for two-phase and about 2.3 db for four-phase modulation.[8,9] Comparable penalties in signal-to-noise ratio as compared to ideal synchronous detection may be incurred with the other methods of providing a demodulating wave mentioned in Section 3.1, owing to small unavoidable amplitude and phase fluctuations in the demodulating wave resulting from other causes than noise. However, in the case of differential phase detection, greater transmission impairments would be expected from phase distortion, since the effect of phase distortion, like that of noise, is present in both the signal and the demodulating wave. The transmission impairments resulting from quadratic delay distortion are determined here, and compared with that encountered with ideal synchronous detection.

Other implementations of differential phase detection than assumed herein have been used, but in principle these entail a wider channel band than with ideal synchronous detection. For example, the two demodulator inputs or outputs could be integrated over a pulse integral T with the aid of a narrow-band resonator tuned to the carrier frequency, and then be rapidly quenched before the next signal interval. When the channel bandwidth is limited, the phase of the demodulating carrier will then depend on the phases of the carrier during several pulse intervals. Thus some intersymbol interference from bandwidth limitation is encountered even in the absence of phase distortion, and the effect of phase distortion will be greater than that determined herein. However, exces-

sive transmission impairments from bandwidth limitation and phase distortion can be avoided by appropriate techniques, as when a large number of narrow channels are provided within a common band of much greater bandwidth than that of the individual channels.[10]

4.2 *Basic Expressions*

In differential phase modulation the carrier would be at midband, i.e., $\omega_c = \omega_0$. With $U_{01} = V$, expression (33) for the demodulated signal becomes

$$V = \sum_{n=-\infty}^{\infty} S_n(0)\bar{P}_0(-n), \tag{96}$$

where

$$S_n(0) = \sum_{m=-\infty}^{\infty} \bar{P}_0(-m) \cos [\psi_0(-n) - \psi_0(-m+1) + \varphi_0(-n) - \varphi_0(-m) - \theta]. \tag{97}$$

The above expressions apply for the output of the single demodulator in two-phase systems, in which $\theta = 0$. In four-phase systems the output of one demodulator is obtained with $\theta = \theta_1$ and the output of the other with $\theta = \theta_2 = \theta_1 \pm \pi/2$.

Examination of (97) shows that the term for $m = n + 1$ is independent of the phase difference $\psi_0(-n) - \psi_0(-m+1)$ and is given by

$$\cos [\varphi_0(-n) - \varphi_0(-n-1) - \theta]\bar{P}_0(-n-1).$$

This represents a DC or bias component. The total bias component in (96) is

$$V_0 = \sum_{n=-\infty}^{\infty} \cos [\varphi_0(-n) - \varphi_0(-n-1) - \theta]\bar{P}_0(-n)\bar{P}_0(-n-1). \tag{98}$$

Determination of transmission impairments becomes rather difficult except for the special case in which $\varphi_0(n) = 0$, which will be considered in further detail below.

4.3 *Even Symmetry Spectrum and Delay Distortion*

When the pulse spectrum has even symmetry about the midband frequency, and the phase characteristic has odd symmetry (i.e., even symmetry delay distortion) the quadrature component of $P_0(t)$ vanishes, i.e., $\varphi(-n) = 0$. In this case, $\bar{P}_0 = R_0$ and (96) becomes

$$\begin{aligned} V &= \sum_{n=-\infty}^{\infty} S_n(0)R_0(-n) \\ &= S_0R_0(0) + \sum_{n=1}^{\infty} [S_nR_0(-n) + S_{-n}R_0(n)], \end{aligned} \tag{99}$$

while (97) and (98) simplify to

$$S_n = S_n(0) = \sum_{m=-\infty}^{\infty} R_0(-m) \cos [\psi_0(-n) - \psi_0(-m+1) - \theta], \tag{100}$$

$$V_0 = \sum_{m=-\infty}^{\infty} R_0(-n)R_0(-n-1) \cos \theta. \tag{101}$$

With $\theta = \pi/4$ in (100) and (101), these expressions can be written, after introduction of a normalizing factor $\sqrt{2}$,

$$S_n = \sum_{m=-\infty}^{\infty} R_0(-m)a_m(n), \tag{102}$$

$$V_0 = \sum_{m=-\infty}^{\infty} R_0(-n)R_0(-n-1), \tag{103}$$

$$\begin{aligned} a_m(n) &= \sqrt{2} \cos [\psi_0(-n) - \psi_0(-m+1) - \pi/4] \\ &= \cos [\psi_0(-n) - \psi_0(-m+1)] + \sin [\psi_0(-n) \\ &\quad - \psi_0(-m+1)]. \end{aligned} \tag{104}$$

With $\psi_0(-n) - \psi_0(-m+1) = 0, \pi/2, \pi$ or $3\pi/2$ the following relations apply

$$\begin{aligned} a_m(n) &= \pm 1 \qquad \text{for} \quad m \neq n+1, \\ &= 1 \qquad\quad \text{for} \quad m = n+1. \end{aligned} \tag{105}$$

In view of (105), (102) can alternately be written in the form

$$S_n - [1 - a_{n+1}(n)]R_0(-n-1) = \sum_{m=-\infty}^{\infty} a_m(0)R_0(-m), \tag{106}$$

where in the summation $a_m(n)$ can be chosen as -1 or as 1, also for $m = n+1$.

4.4 *Two-Phase Modulation*

With synchronous detection, two-phase modulation could be used in conjunction with both double-sideband and vestigial-sideband transmission. With differential phase detection, however, vestigial-sideband

transmission is not practicable, since severe transmission impairments would be incurred even in the absence of phase distortion, owing to the presence of the quadrature component. Hence only double-sideband two-phase modulation is considered here.

For $n = 0$, (106) gives

$$S_0 - [1 - a_1(0)R_0(-1)] = \sum_{m=-\infty}^{\infty} a_m(0)R_0(-m). \tag{107}$$

Assume that in (107) a sequence of values of $a_m(0)$ has been chosen, for example $a_{-3}(0) = 1$, $a_{-2}(0) = 1$, $a_{-1}(0) = 1$, $a_0(0) = 1$, $a_1(0) = -1$, $a_2(0) = 1$, etc. For any other value of n than $n = 0$, the sequence of $a_m(n)$ will either be identical with that for $a_m(0)$, or all signs will be reversed. This follows from (104), since $\psi_0(-n)$ will differ from $\psi_0(0)$ by 0 or π. Hence, for $n \neq 0$, the right-hand side of (106) can be replaced by the left-hand side of (107), so that

$$S_n - [1 - a_{n+1}(n)]R_0(-n-1) = \pm [S_0 - [1 - a_1(0)]R_0(-1)] \tag{108}$$

In the absence of intersymbol interference, V as given by (99) would be -1 or 1. In the following, the minimum possible value of V will be determined, on the assumption that $V = 1$ without intersymbol interference; i.e., $a_0(0) = 1$.

Consider first the term $S_0R_0(0)$ in (99). The minimum possible value of S_0 is obtained from (107) by choosing $a_m(0) = -1$ for $R_0(-m) > 0$ and $a_m(0) = 1$ for $R_0(-m) < 0$. The following relation is thus obtained for the minimum possible value of S_0, on the above premise of $a_0(0) = 1$:

$$\begin{aligned} S_{0,\min} = {} & [1 - a_1(0)]\, R_0(-1) + R_0(0) \\ & - \sum_{m=1}^{\infty} [\bar{R}_0(-m) + \bar{R}_0(m)], \end{aligned} \tag{109}$$

where $\bar{R}_0$ designates the absolute values. In the above expression $a_1(0)$ would be taken as $a_1(0) = 1$ if $R_0(-1) < 0$ and as $a_1(0) = -1$ if $R_0(-1) > 0$. The term $[1 - a_1(0)]R_0(-1)$ can therefore be written alternatively as $R_0(-1) + \bar{R}_0(-1)$, in which case (109) becomes

$$S_{0,\min} = R_0(-1) + \bar{R}_0(-1) + U_{\min}, \tag{110}$$

where

$$U_{\min} = R_0(0) - \sum_{m=1}^{\infty} [\bar{R}_0(-m) + \bar{R}_0(m)]. \tag{111}$$

It will be recognized that $U_{\min}$ is the same as the minimum possible value

of the demodulated voltage with synchronous detection, in the presence of a mark, as given in a somewhat more general form by (73).

Having thus determined $S_{0,\min}$ it follows from (108) and (110) that the two possible associated values of $S_{n,\min}$ are given by

$$S_{n,\min} = R_0(-n-1) + \bar{R}_0(-n-1) \pm U_{\min}, \tag{112}$$

where the term $[1 - a_{n+1}(n)]R_0(-n-1)$ in (108) has been replaced by the equivalent representation by the first two terms in (112).

To obtain the minimum value of V as given by (99), each term in the series must be made to have the maximum negative value. To this end the negative sign in (112) for $U_{\min}$ is chosen if $R_0(n)$ is positive, and the positive sign if $R_0(n)$ is negative. The minimum possible value of V thus obtained with (110) and (112) in (99) is

$$\begin{aligned} V_{\min} = {} & [R_0(-1) + \bar{R}_0(-1) + U_{\min}]R_0(0) \\ & + \sum_{n=1}^{\infty} [R_0(-n-1) + \bar{R}_0(-n-1)]R_0(-n) \\ & + \sum_{n=1}^{\infty} [R_0(n+1) + \bar{R}_0(n+1)]R_0(n) \\ & - U_{\min} \sum_{n=1}^{\infty} [\bar{R}_0(-n) + \bar{R}_0(n)], \end{aligned} \tag{113}$$

$$\begin{aligned} = {} & U_{\min}\left\{R_0(0) - \sum_{n=1}^{\infty} [\bar{R}_0(-n) + \bar{R}_0(n)]\right\} \\ & + \sum_{n=-\infty}^{\infty} [R_0(-n-1) + \bar{R}_0(-n-1)]R_0(-n), \end{aligned} \tag{114}$$

where the first term can also be written ${U_{\min}}^2$.

In accordance with the discussion in Section 4.2, the demodulator output contains a bias or DC component V_0 given by (103). Optimum performance is obtained when the threshold level for distinction between $V = 1$ and $V = -1$ is made equal to V_0. When V_0 is subtracted from both sides of (114) the following expression is obtained for two-phase modulation:

$${V_{\min}}^0 = V_{\min} - V_0 = {U_{\min}}^2 + \Sigma, \tag{115}$$

where

$$\Sigma = \sum_{n=-\infty}^{\infty} \bar{R}_0(-n-1)R_0(-n). \tag{116}$$

When intersymbol interference is absent at sampling instants, $R_0(n) = 0$ for $n \neq 0$, and for $n = 0$ is $R_0^0(0)$. In this case $V_{\min}^0 = U_{\min}^2 = [R_0^0(0)]^2$. The voltage given by (115) is smaller than in the absence of intersymbol interference by the factor

$$\eta_{\min}^0 = \eta_{\min}^2 + \Sigma/[R_0^0(0)]^2, \tag{117}$$

where $\eta_{\min}$ applies for synchronous detection.

4.5 *Four-Phase Modulation*

The basic difference between two-phase and four-phase modulation is that relation (108) does not apply for four-phase modulation. Returning to the discussion following (107), if a sequence $a_m(0)$ is chosen in four-phase transmission, the sequence $a_m(n)$ can be chosen independently. This follows from the (104), which shows that if $a_m(0)$ has a given value, say $a_m(0) = 1$, it is possible to make each $a_m(n)$ equal to $+1$ or -1 by appropriate choice of $\psi(-n)$.

For this reason the minimum value (or maximum negative value) of the right-hand side of (106) is now, for $n \neq 0$:

$$[S_n - [1 - a_{n+1}(n)]R_0(-n-1)]_{\min} = -R_0(0) - \sum_{m=1}^{\infty} [\bar{R}_0(-m) + \bar{R}_0(m)]. \tag{118}$$

The right-hand side of (118) is smaller than for two-phase transmission as given by (112) by $-2R_0(0)$. When this modification is introduced, the following expression is obtained for four-phase modulation, in place of (115) for two-phase modulation:

$$V_{\min}^0 = U_{\min}^2 - 2R_0(0) \sum_{n=1}^{\infty} [\bar{R}_0(-n) + \bar{R}_0(n)] + \sum_{n=-\infty} \bar{R}_0(-n-1)R_0(-n) \tag{119}$$

or

$$V_{\min}^0 = U_{\min}^2 - 2R_0(0)[R_0(0) - U_{\min}] + \Sigma, \tag{120}$$

where Σ is given by (116).

The voltage given by (120) is smaller than in the absence of intersymbol interference by the factor

$$\eta_{\min}^0 = [V_{\min}^0/R_0^0(0)]^2. \tag{121}$$

TABLE X — MINIMUM AMPLITUDES OF DEMODULATED PULSE TRAINS IN TWO-PHASE MODULATION WITH DIFFERENTIAL PHASE DETECTION

d/T	0	1	2	3	4
$U_{\min}^2$	1	0.730	0.536	0.273	0.120
Σ	0	0.012	0.058	0.137	0.222
$V_0 = L_0$	0	0.012	0.058	0.137	0.222
$V_{\min}^0 = \eta_{\min}^0$	± 1	± 0.74	± 0.594	± 0.41	± 0.342
$V_{\min}^+ = V_{\min}^0 + V_0$	1	0.75	0.652	0.547	0.564
$V_{\min}^- = -V_{\min}^0 + V_0$	-1	-0.73	-0.536	-0.273	-0.12

4.6 *Raised Cosine Spectrum and Quadratic Delay Distortion*

The function $R_0(n)$ for this case is given in Table III. The values of $U_{\min}$ for synchronous detection are given in Table IV for $l = 2$. In Table X are given the various quantities appearing in expression (115) for the minimum amplitudes of received pulse trains at sampling instants with optimum slicing lead equal to the DC component V_0. The values of $V_{\min}^0 = \eta_{\min}^0$ are shown in Fig. 8.

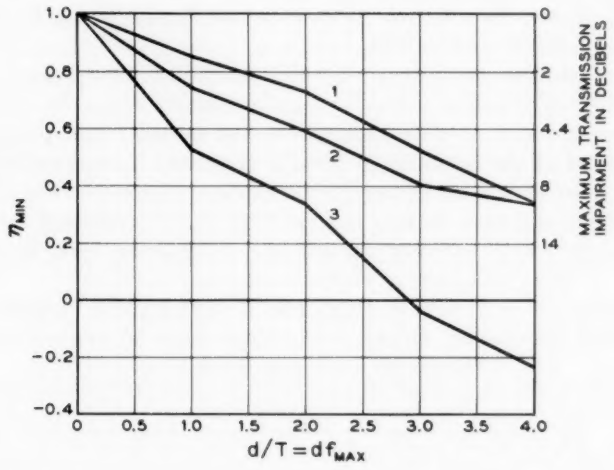


Fig. 8 — Factor $\eta_{\min}$ for raised cosine spectrum and quadratic delay distortion as in Fig. 2 for synchronous detection and differential phase detection. Curve 1: Ideal synchronous detection — applies for two-phase and four-phase modulation with carrier at midband and pulses at intervals T, and for vestigial-sideband transmission with pulses at intervals $T/2$. Curve 2: Ideal differential phase detection — two-phase modulation with pulse interval T. Curve 3: Ideal differential phase detection — four-phase modulation with pulse interval T.

TABLE XI — MINIMUM AMPLITUDES OF DEMODULATED PULSE TRAINS IN FOUR-PHASE MODULATION WITH DIFFERENTIAL PHASE DETECTION

d/T	0	1	2	3*	4*
$V_0 = L_0$	0	0.012	0.058	−0.137	−0.232
$V_{\min}{}^0 = \eta_{\min}{}^0$	±1	±0.53	±0.33	±0.04	±0.23
$V_{\min}{}^+ = V_{\min}{}^0 + V_0$	+1	+0.54	+0.39	+0.10	0
$V_{\min}{}^- = -V_{\min}{}^0 + V_0$	−1	−0.52	−0.27	+0.18	+0.45

* Reversal of sign indicates a reversal in sign of the demodulated pulses.

With the accuracy used herein it turns out that Σ and V_0 are numerically equal but are not identical.

It will be noted that, when delay distortion is pronounced, the bias component V_0 is appreciable, and that a significant penalty can be incurred if the threshold or slicing level is taken as 0 rather than V_0. For example, with $d/T = 4$ and 0 threshold level the minimum amplitude of a demodulated pulse for a carrier phase $\psi = 0$ would be 0.564, and the minimum negative amplitude for a carrier phase $\psi = \pi$ would be −0.12. With the optimum threshold level the minimum amplitudes are ±0.342. Hence the tolerable peak noise amplitudes would be greater by a factor $0.342/0.12 = 2.85$.

With four-phase modulation the values given in Table XI are obtained from (120). The values of $V_{\min}{}^0 = \eta_{\min}{}^0$ are shown in Fig. 8.

In the above illustrative examples it was assumed that pulses were transmitted at the minimum interval T permitted if intersymbol interference is to be avoided in the absence of delay distortion. The effect of delay distortion may or may not be reduced by increasing the pulse interval, that is, in exchange for a slower transmission rate. By way of

TABLE XII — FUNCTION $R_0(n)$ FOR RAISED COSINE SPECTRUM AND QUADRATIC DELAY DISTORTION WITH 50 PER CENT INCREASE IN PULSE INTERVAL

n^*	d/T				
	0	1	2	3	4
−2	0	−0.0006	0.0025	0	[illegible]
−1	0	0.0053	0.0081	0.0145	0.0202
0	1	0.9633	0.8795	0.7956	0.7336
1	0	−0.0341	−0.1161	−0.2200	−0.2909
2	0	0.0044	0.0020	−0.0231	−0.0584
3	0	−0.0009	0.0011	0.0044	−0.0030

$n^* = \frac{2}{3}$ of the values n given in Table XIX.

TABLE XIII — MINIMUM AMPLITUDES OF DEMODULATED PULSES WITH 50 PER CENT INCREASE IN PULSE INTERVALS

d/T	0	1	2	3	4
U_{min}	1	0.92	0.75	0.50	0.36
Σ	0	−0.028	−0.10	−0.17	−0.08
$V_{min}{}^0$ (2-phase)	1	0.80	0.46	0.08	0.045
$V_{min}{}^0$ (4-phase)	1	0.72	0.23	−0.40	−0.50

illustration it will be assumed that the pulse interval is increased by a factor 1.5, in which case the values of R_0 are as given in Table XII.

With this modification, the various quantities are as given in Table XIII.

In Fig. 9 values of U_{min} and $V_{min}{}^0$ are compared with those for the minimum interval between pulses. It will be noted that there is no significant difference in the case of two-phase or four-phase modulation with synchronous detection. With differential synchronous detection some advantage is realized for small delay distortion in exchange for a disadvantage with pronounced delay distortion.

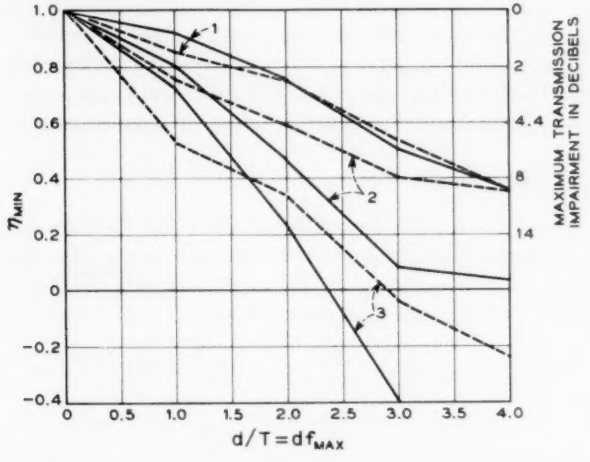


Fig. 9 — Effect of pulse interval on factor η_{min} for raised cosine spectrum with quadratic delay distortion (dashed curves: pulse interval T, as in Fig. 8; solid curves: pulse interval $1.5T$). Curves 1: Ideal synchronous detection — applies for two-phase and four-phase modulation with carrier at midband and pulses at intervals $1.5T$ and for vestigial-sideband transmission with pulses at intervals $0.75T$. Curves 2: Ideal differential phase detection — two-phase modulation with pulse intervals $1.5T$. Curves 3: Ideal differential phase detection — four-phase modulation with pulse intervals $1.5T$.

V. BINARY FREQUENCY MODULATION (FSK)

5.1 *General*

As shown elsewhere,[1] with optimum design, binary FM or frequency shift keying requires the same minimum bandwidth as double-sideband AM. In the absence of transmission distortion from gain and phase deviations, the optimum signal-to-noise ratio required at the detector input for a given error probability is slightly greater than for two-phase transmission with ideal synchronous detection, but would be expected to be about the same as for two-phase transmission with ideal differential phase detection. Binary FM may be preferable to the latter method from the standpoint of implementation and has an advantage over the simpler method of binary AM with envelope detection from the standpoint of signal-to-noise ratio and performance during sudden transmission level variations.

The performance of binary FM is determined here for channels with linear and quadratic delay distortion and compared with that of the other methods mentioned above. In this analysis ideal frequency discriminator detection is assumed, in which the demodulated signal is proportional to the time derivative of the phase of the received wave. This condition may be closely approached with actual detectors when the channel bandwidth is small in relation to the midband frequency. However, when this is not the case, ideal FM detection is only approximated with conventional frequency discriminators or zero-crossing detectors.

5.2 *Basic Expression*

Expression (48) for the demodulated pulse train applies for any amplitude and phase characteristic of the channel. In the case of a continuing space, $a(n) = 0$ in (46) and (47) and $U_0(t) = 0$. With a continuing mark, $a(n) = 1$ in the above expressions and

$$\alpha_0(x) = \sum_{n=-\infty}^{\infty} (-1)^n R_0(x - n), \tag{122}$$

$$\beta_0(x) = \sum_{n=-\infty}^{\infty} (-1)^n Q_0(t - n). \tag{123}$$

Returning to (21), it will be recognized that (122) and (123) represent the in-phase and quadrature components in a binary amplitude modulation system when pulses of duration T and alternating polarity are transmitted, i.e., $a(n) = (-1)^n$ in (21). The fundamental frequency of such a pulse train is $\tilde{\omega} = T/\pi$. Let $A(-\tilde{\omega})$ and $\Psi(-\tilde{\omega})$ be the ampli-

tude and phase characteristic of the channel at the frequency $-\bar{\omega}$ from ω_0; $A(\bar{\omega})$ and $\Psi(\bar{\omega})$ the corresponding quantities at the frequency $\bar{\omega}$ from ω_0. Solution of (122) and (123) for the above steady state condition of alternate marks and spaces in binary AM gives

$$\alpha_0(x) = A(-\bar{\omega}) \cos [-\bar{\omega}t - \Psi(-\bar{\omega})] + A(\bar{\omega}) \cos [\bar{\omega}t - \Psi(\bar{\omega})], \quad (124)$$

$$\beta_0(x) = -A(\bar{\omega}) \sin [-\bar{\omega}t - \Psi(-\bar{\omega})] + A(\bar{\omega}) \sin [\bar{\omega}t - \Psi(\bar{\omega})]. \quad (125)$$

With (124) and (125) in (50), it turns out by way of check that $U_0(x) = 1$ for a continuing mark for any amplitude and phase characteristic of the channel.

For pulse trains other than continuing marks or spaces, intersymbol interference will be encountered from amplitude and phase distortion. In the following section special cases of phase distortion will be examined further. It will be assumed that the amplitude characteristic has the appropriate shape so that intersymbol interference can be avoided in the absence of phase distortion. To this end it is necessary that $A(-\bar{\omega}) = A(\bar{\omega}) = \frac{1}{2}$ or $\mu = 2$ as shown elsewhere (Ref. 1, Section V). In this case (50) becomes with $U_0 = U$:

$$U(t) = \frac{1}{D}\Big[2(\alpha_0^2 + \beta_0^2) - \alpha_0 \cos y - \beta_0 \sin y - \frac{1}{\bar{\omega}} (\alpha_0' \sin y - \beta_0' \cos y) - \frac{2}{\bar{\omega}} (\beta_0'\alpha_0 - \alpha_0'\beta_0)\Big], \quad (126)$$

where

$$D = 1 + 4(\alpha_0^2 + \beta_0^2) - 4(\alpha_0 \cos y + \beta_0 \sin y), \quad (127)$$

in which

$$\alpha_0' = d\alpha_0/dt,$$
$$\beta_0' = d\beta_0/dt, \quad (128)$$
$$y = \pi x + \Psi(-\bar{\omega}),$$

$$\alpha_0 = \alpha_0(x) = \sum_{n=-\infty}^{\infty} (-1)^n a(n) R_0(x - n), \quad (129)$$

$$\beta_0 = \beta_0(x) = \sum_{n=-\infty}^{\infty} (-1)^n a(n) Q_0(x - n), \quad (130)$$

where

$$x = t/T, \qquad a(n) = \begin{matrix} 0 \\ 1 \end{matrix} \begin{matrix} \text{for space} \\ \text{for mark.} \end{matrix}$$

5.3 *Even Symmetry Spectrum and Delay Distortion*

When the amplitude characteristic of the channel has even symmetry about ω_0 and the phase characteristic has odd symmetry $Q_0(x + n) = 0$ and (126) simplifies to

$$U(t) = \frac{2\alpha_0^2 - \alpha_0 \cos y - (\alpha_0'/\tilde{\omega}) \sin y}{1 + 4\alpha_0^2 - 4\alpha_0 \cos y}. \tag{131}$$

Optimum performance would be expected when a single pulse is sampled at its peak, a condition which is at least closely approximated with $y = 0$. This condition is met when $t = t_0$ is so chosen that

$$t_0/T = x_0 = -\Psi(-\tilde{\omega})/\pi. \tag{132}$$

Expression (131) then simplifies to

$$U(t_0) = \frac{\alpha_0(x_0)}{2\alpha_0(x_0) - 1}. \tag{133}$$

For further analysis it is convenient to introduce the quantities

$$\alpha_0^-(x_0) = \sum_{n=1}^{\infty} |(-1)^n R_0(x - n)|^- + |(-1)^n R_0(x + n)|^-, \tag{134}$$

$$\alpha_0^+(x) = \sum_{n=1}^{\infty} |(-1)^n R_0(x - n)|^+ + |(-1)^n R_0(x - n)|^-, \tag{135}$$

where $|\ \ |^-$ designates absolute values when $(-1)^n R_0(x \pm n)$ is negative and $|\ \ |^+$ when it is positive.

It will be recognized that

$$\begin{aligned} R_0(x) + \alpha_0^+(x) - \alpha_0^-(x) &= \sum_{n=-\infty}^{\infty} (-1)^n R_0(x - n) \\ &= \alpha_0(x) = \cos y, \end{aligned} \tag{136}$$

where the last relations follow in view of (122), (124) and (128).

During transmission of a space, delay distortion will have an adverse effect only if U as given by (133) is positive, since only in this case is the tolerance to noise reduced. To obtain a positive value of U, it is necessary to have either $\alpha_0 \geqq \frac{1}{2}$ or $\alpha_0 < 0$. For a space, $a(0) = 0$ in (129) and a value of $\alpha_0 \geqq \frac{1}{2}$ can then be excluded for any reasonable delay distortion. It will, therefore, be assumed that $\alpha_0 < 0$. The maximum positive value of U, i.e., the maximum adverse effect of delay distortion, is then obtained with the maximum possible negative value of α_0. This maximum value is obtained by choosing $a(n) = 0$ in (129)

whenever $(-1)^n R_0(x-n)$ is positive and choosing $a(n) = 1$ whenever $(-1)^n R_0(x-n)$ is negative. The maximum negative value of $\alpha_0(x)$ thus obtained is given by (134). The corresponding maximum value of $U(t_0)$ in the presence of a space and with sampling at $t = t_0$ as defined by (132) is

$$U_{\max}^{(0)} = \frac{-\alpha_0^-(x_0)}{-2\alpha_0^-(x_0) - 1} = \frac{\alpha_0^-(x_0)}{1 + 2\alpha_0^-(x_0)}. \tag{137}$$

During transmission of a mark delay distortion will have an adverse effect only if $U(t_0) < 1$. This will be the case if $\alpha_0 > 1$ or $\alpha_0 < \frac{1}{2}$ in (133). With $a(0) = 1$ in (129) for a mark, the condition $\alpha_0 < \frac{1}{2}$ will not be encountered with any reasonable delay distortion and only the case $\alpha_0 > 1$ needs to be considered. The minimum positive value of $U(t_0)$ in the presence of a mark is obtained when α_0 is taken as the maximum positive value given by $\alpha_0(x) = R_0(x) + \alpha_0^+(x)$, where α_0^+ is given by (135). In view of (136) it follows that, for $y = 0$,

$$\alpha_0(x_0) = R_0(x_0) + \alpha_0^+(x_0) = 1 + \alpha_0^-(x_0). \tag{138}$$

With (138) in (133) the minimum amplitude of a pulse train in the presence of a mark and at the sampling instant t_0 defined by (132) becomes:

$$U_{\min}^{(1)} = \frac{1 + \alpha_0^-(x_0)}{2[1 + \alpha_0^-(x_0)] - 1} = \frac{1 + \alpha_0^-(x_0)}{1 + 2\alpha_0^-(x_0)}. \tag{139}$$

The optimum slicing level in the presence of delay distortion becomes for conditions as discussed in Section 2.9,

$$L_0 = \tfrac{1}{2}[U_{\max}^{(0)} + U_{\min}^{(1)}] = \tfrac{1}{2}. \tag{140}$$

The minimum amplitude of a pulse train in the presence of a mark relative to the optimum slicing level becomes

$$U_{\min}^{(1)} - L_0 = \frac{1}{2}\frac{1}{[1 + 2\alpha_0^-(x_0)]}. \tag{141}$$

The latter expression also applies for the difference between the slicing level and the maximum amplitude of a pulse train at a sampling point in the presence of a space.

Expression (141) shows that the minimum amplitude at a sampling point is smaller than in the absence of delay distortion ($\alpha_0^- = 0$) by the factor

$$\eta_{\min} = \frac{1}{1 + 2\alpha_0^-(x_0)}. \tag{142}$$

TABLE XIV — FACTOR η_{min} FOR RAISED COSINE SPECTRUM AND QUADRATIC DELAY DISTORTION

d/T	0	1	2	3	4
$\Psi(-\tilde{\omega})$	0	$-\pi/12$	$-\pi/6$	$-\pi/4$	$-\pi/3$
x_0	0	1/12	1/6	1/4	1/3
$\alpha_0^-(x_0)$	0	0.07	0.125	0.20	0.35
η_{min}	1	0.88	0.80	0.72	0.59

5.4 *Raised Cosine Spectrum and Quadratic Delay Distortion*

In the particular case of a raised cosine spectrum of the pulses at the detector input, as shown in Fig. 1, and quadratic delay distortion, the function $R_0(t/T) = R_0(x + n)$ is given in Table XIX of the Appendix. The phase distortion $\Psi(-\tilde{\omega})$ in this case is given by

$$\Psi(-\tilde{\omega}) = -\left(\frac{d}{T}\right)\frac{\pi}{12}, \tag{143}$$

where d/T is defined as in Section 4.5.

In Table XIV are given $\Psi(-\tilde{\omega})$ together with x_0 as obtained from (132), $\alpha_0^-(x_0)$ as given by (134) and η_{min} as obtained from (142). These factors are shown in Fig. 10, together with the corresponding factor for binary double-sideband AM as obtained from Table IV.

5.5 *Raised Cosine Spectrum and Linear Delay Distortion*

When both the pulse spectrum at the detector input and phase distortion has even symmetry about the frequency ω_0, the following relations apply (see Appendix):

$$R_0(-t) = R_0(t), \qquad Q_0(-t) = Q_0(t); \tag{144}$$

$$R_0'(-t) = -R_0'(t), \qquad Q_0'(-t) = -Q_0'(t). \tag{145}$$

The maximum amplitude of a single pulse in this case is at $t = 0$. Optimum performance is obtained with sampling at $t = 0$, in which case y in (126) and (127) is given by

$$y = y_0 = \Psi(-\tilde{\omega}), \tag{146}$$

and:

$$\alpha_0 = \alpha_0(0) = \sum_{n=-\infty}^{\infty} (-1)^n a(n) R_0(-n), \tag{147}$$

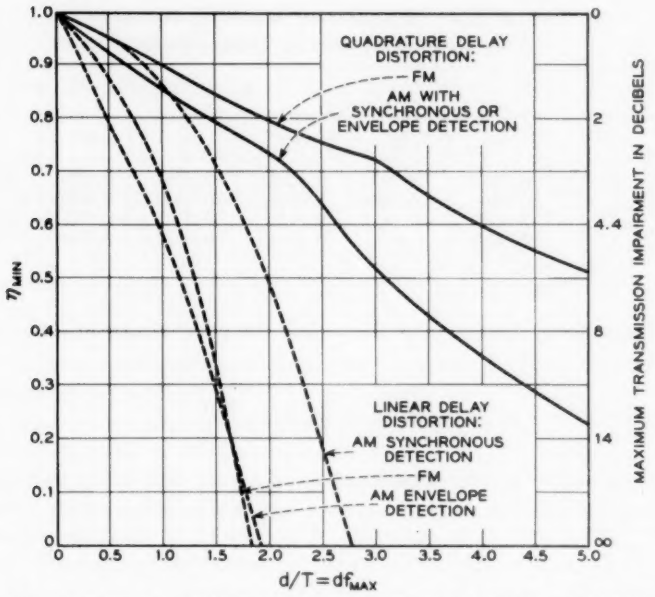


Fig. 10 — Factor η_{min} for binary pulse transmission by FM and double-sideband AM, for quadratic delay distortion as in Fig. 2 and linear delay distortion as in Fig. 4.

$$\beta_0 = \beta_0(0) = \sum_{n=-\infty}^{\infty} (-1)^n a(n) Q_0(-n), \tag{148}$$

$$\alpha_0' = \alpha_0'(0) = \sum_{n=-\infty}^{\infty} (-1)^n a(n) R_0'(n), \tag{149}$$

$$\beta_0' = \beta_0'(0) = \sum_{n=-\infty}^{\infty} (-1)^n a(n) Q_0'(n). \tag{150}$$

For the special case of a raised cosine spectrum and linear delay distortion, the functions R_0 and Q_0 are given in Table XX of the Appendix. The functions $R_0'(n)$ and $Q_0'(n)$ are related to the functions R_1 and Q_1 given in Table XXII by

$$R_0'(t)/\bar{\omega} = \frac{4}{\pi} R_1(t), \qquad Q_0'(t)/\bar{\omega} = \frac{4}{\pi} Q_1(t). \tag{151}$$

The functions R_0 and Q_0 are given in Table XV for integral values of $n = t/T$ and the functions $R_0'/\bar{\omega}$ and $Q_0'/\bar{\omega}$ are given in Table XVI.

In the above case of quadratic phase distortion of the form $\psi(u) = cu^2$

TABLE XV — FUNCTIONS $R_0(nT)$ AND $Q_0(nT)$ FOR LINEAR DELAY DISTORTION

d/T	0.5		1		2	
n	R_0	Q_0	R_0	Q_0	R_0	Q_0
0	0.9516	0.1941	0.8309	0.3306	0.5786	0.3895
±1	0.080	−0.0956	0.0726	−0.1792	0.2596	−0.2439
±2	0.0048	−0.0014	0.0112	0.0107	−0.0383	0.0577
±3	0.0008	0	0.0006	0.0024	−0.0094	−0.0052
±4	0	0	0	0	0	0

and linear delay distortion $\psi'(u) = 2cu$, the phase distortion at $u = -\bar{\omega}$ is given by

$$\Psi(-\bar{\omega}) = \Psi(\bar{\omega}) = \frac{d}{T}\frac{\pi}{4}. \tag{152}$$

Owing to the several quantities α_0, β_0, α_0', β_0', $\beta_0'\alpha_0$, $\alpha_0'\beta_0$ and y_0 involved in (126), it does not appear feasible to derive simple relations for $U_{\max}^{(s)}$ and $U_{\min}^{(m)}$. However, it is possible to determine these by examining several cases, as illustrated below for $d/T = 1$ and $d/T = 2$.

With $d/T = 1$ in (152), relation (146) gives

$$y_0 = \Psi(-\bar{\omega}) = \pi/4$$

and (126) becomes

$$U(0) = N_1/D_1, \tag{153}$$

$$N_1 = 2(\alpha_0^2 + \beta_0^2) - \tfrac{1}{2}\sqrt{2}(\alpha_0 + \beta_0 + \alpha_0'/\bar{\omega} - \beta_0'/\bar{\omega}) - 2[(\alpha_0\beta_0'/\bar{\omega} - \beta_0\alpha_0'/\bar{\omega})], \tag{154}$$

$$D_1 = 1 + 4(\alpha_0^2 + \beta_0^2) - 2\sqrt{2}(\alpha_0 + \beta_0), \tag{155}$$

where α_0, β_0, α_0' and β_0' are given by (147) through (150).

TABLE XVI — FUNCTIONS $R_0'/\bar{\omega} = (4/\pi)R_1$ AND $Q_0'/\bar{\omega} = (4/\pi)Q_1$ FOR LINEAR DELAY DISTORTION

d/T	0.5		1		2	
n	$R_0'/\bar{\omega}$	$Q_0'/\bar{\omega}$	$R_0'/\bar{\omega}$	$Q_0'/\bar{\omega}$	$R_0'/\bar{\omega}$	$Q_0'/\bar{\omega}$
0	0	0	0	0	0	0
±1	±0.1432	±0.0624	±0.2263	±0.0639	±0.3016	±0.1053
±2	±0.0065	±0.0134	±0.0013	±0.0272	±0.0980	±0.0167
±3	±0.0006	±0.0036	±0.0051	±0.0023	±0.0042	±0.0188
±4	0	0	0	0	0	0

TABLE XVII — VALUES OF $U(0)$ FOR RAISED COSINE SPECTRUM AND LINEAR DELAY DISTORTION WITH $d/T = 1$

$a(-2)$	$a(-1)$	$a(0)$	$a(1)$	$a(2)$	α_0	β_0	$\alpha_0'/\bar{\omega}$	$\beta_0'/\bar{\omega}$	N_1	D_1	$U(0)$
0	1	0	1	0	−0.146	0.358	0	0	0.15	1.0	0.15
0	0	0	1	0	−0.073	0.179	−0.227	0.064	0.173	0.87	0.16*
0	1	0	0	0	−0.073	0.179	0.227	−0.064	−0.29	0.87	−0.33
1	1	0	1	1	−0.124	0.38	0	0	0.136	0.90	0.15
0	0	1	0	0	0.831	0.331	0	0	0.78	0.90	0.86
0	1	1	1	0	0.68	0.89	0	0	1.4	1.6	0.87
0	0	1	1	0	0.76	0.51	−0.227	0.064	0.66	0.76	0.87
0	1	1	0	0	0.76	0.51	0.227	−0.064	0.89	0.76	1.16
1	1	1	1	1	0.70	0.91	0	0	1.21	1.43	0.84*

For various combinations of marks and spaces, i.e., $a(n) = 1$ and 0, the results given in Table XVII are obtained.

The maximum value in the presence of a space is $U_{\max}^{(0)} \cong 0.16$ and the minimum value in the presence of a mark is $U_{\min}^{(1)} \cong 0.84$, as indicated by asterisks. The optimum slicing level is $\frac{1}{2}[U_{\max}^{(0)} + U_{\min}^{(1)}] = 0.5$. The factor $\eta_{\min}$ is in this case

$$\eta_{\min} = U_{\min}^{(1)} - U_{\max}^{(0)} \cong 0.68.$$

For $d/T = 2$, (152) gives $\Psi(-\bar{\omega}) = \pi/2 = y_0$. In this case (126) becomes

$$U(0) = N_2/D_2, \tag{156}$$

$$N_2 = 2(\alpha_0^2 + \beta_0^2) - \beta_0 - \alpha_0'/\bar{\omega} - 2(\alpha_0\beta_0'/\bar{\omega} - \beta_0\alpha_0'/\bar{\omega}), \tag{157}$$

$$D_2 = 1 + 4(\alpha_0^2 + \beta_0^2) - 4\beta_0. \tag{158}$$

The results in this case are given in Table XVIII. In this case $U_{\max}^{(0)} = 0.57$, $U_{\min}^{(1)} = 0.42$, $L_0 \cong 0.5$ and $\eta_{\min} = U_{\min}^{(1)} - U_{\max}^{(0)} \cong -0.15$.

TABLE XVIII — VALUES OF $U(0)$ FOR RAISED COSINE SPECTRUM AND LINEAR DELAY DISTORTION WITH $d/T = 2$

$a(-2)$	$a(-1)$	$a(0)$	$a(1)$	$a(2)$	α_0	β_0	$\alpha_0'/\bar{\omega}$	β_0'/ω	N_2	D_2	$U(0)$
0	1	0	1	0	−0.520	0.488	0	0	0.526	1.08	0.49
0	0	0	1	0	−0.260	0.244	−0.302	−0.106	0.09	0.53	0.17
0	1	1	0	0	−0.260	0.244	0.302	0.106	−0.07	0.53	−0.13
1	1	0	1	1	−0.596	0.603	0.07	0.06	0.837	1.96	0.57*
0	0	1	0	0	0.579	0.390	0	0	0.58	1.38	0.42*
0	1	1	1	0	0.069	0.88	0	0	0.66	0.56	1.17
0	0	1	1	0	0.319	0.634	−0.302	−0.106	0.46	0.67	0.69
0	1	1	0	0	0.319	0.634	0.302	−0.106	0.48	0.67	0.71

The factors in Table XVIII are shown in Fig. 10, together with the corresponding factors for binary double-sideband AM with synchronous detection as given in Table V and with envelope detection as given in Table IX.

VI. SUMMARY

6.1 *General*

The shape of pulse trains at the detector input and output in pulse transmission by various methods of carrier modulation and detection has been formulated in terms of a basic function common to all modulation methods: the carrier pulse transmission characteristic. This function is related to the amplitude and phase characteristics of the channel by a Fourier integral, which can be evaluated by numerical integration with the aid of digital computers for any prescribed channel characteristic. In this way can be determined the effect of specified channel gain and phase deviations on the demodulated pulse train for any modulation method, together with the resultant maximum transmission impairment.

The carrier pulse transmission characteristics are given herein for the representative case of pulses with a raised cosine spectrum at the detector input, for two cases of envelope delay distortion over the channel band. In one case delay distortion is assumed to vary linearly with frequency, and in the other case to vary as the second power of frequency from midband, as indicated in Fig. 11. The resultant maximum effect on the demodulated pulse trains at sampling instants has been determined for various carrier modulation and detection methods, together with the corresponding maximum transmission impairment. The maximum transmission impairment is expressed as the maximum increase in signal-to-noise ratio required at the detector input to compensate for the effect of phase distortion, or corresponding envelope delay distortion. The maximum transmission impairments specified here apply as the error probability approaches zero, and actual impairment will be somewhat smaller, depending on error probability.

In evaluating the effect of phase distortion, idealized modulation and demodulation have been assumed, together with ideal implementation in other respects, such as instantaneous sampling of the appropriate instants and optimum slicing levels.

The numerical results are given in various tables and curves, summarized in Fig. 12 and discussed briefly below.

6.2 *Choice of Transmission Delay Parameters*

In the expressions for the carrier pulse transmission characteristic the phase characteristic of the channel is a basic function. Transmission impairments from phase distortion could be expressed in terms of some parameter or set of parameters that would define the type of phase distortion under consideration. Alternatively, any type of phase distortion can be specified in terms of its derivative with respect to frequency, that is, in terms of envelope delay distortion. From the standpoint of engineering applications the latter method is preferable, since variation in transmission delay over the channel band is more readily measured than variation in phase, and it is ordinarily the quantity specified for various existing facilities.

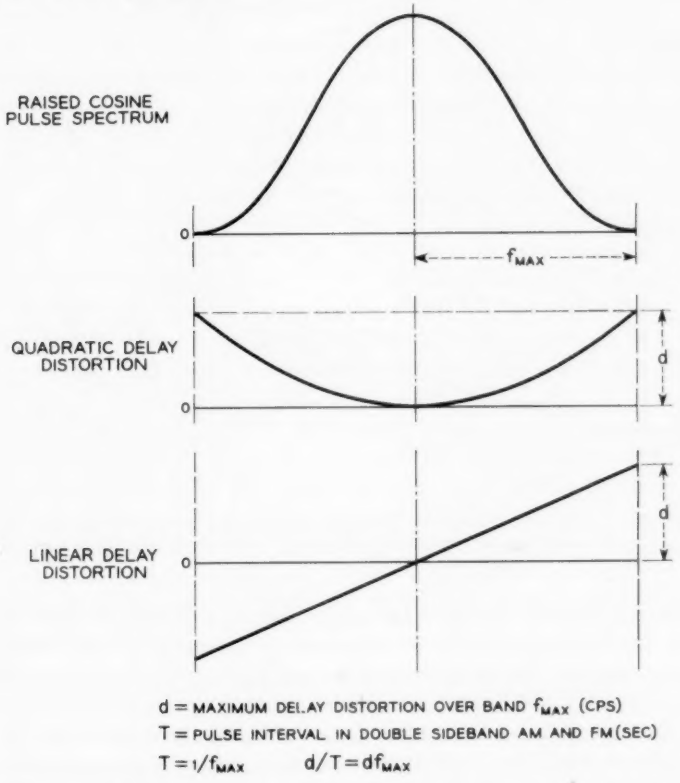


Fig. 11 — Pulse spectrum at detector input and types of delay distortion assumed in comparison of modulation methods.

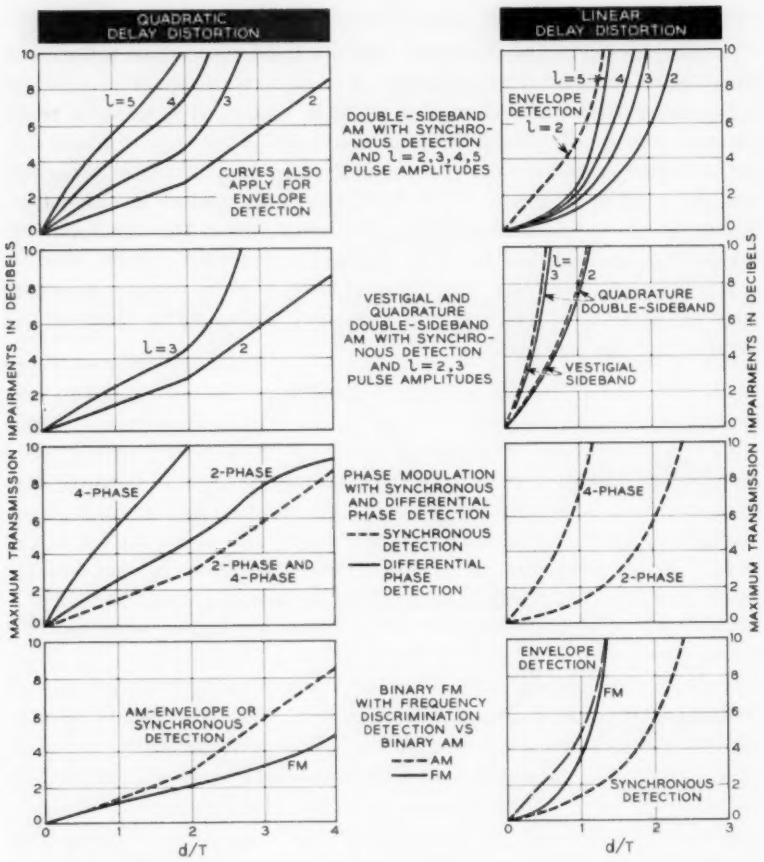


Fig. 12 — Maximum transmission impairments with various modulation methods for raised cosine pulse spectrum with linear and quadratic delay distortion as in Fig. 11.

Linear, quadratic or any other analytically specified delay distortion can be expressed in terms of the difference in transmission delay between any two reference frequencies in the channel band. In the present analysis the difference d in delay between the midband frequency and the maximum frequency f_{max} from midband, as in Fig. 11, has been taken as a basic parameter. The maximum transmission impairments with various carrier modulation methods have been given in terms of the ratio $d/T = df_{max}$, where T is the pulse interval in double-sideband AM.

An alternative choice of delay parameter might have been the maximum

difference d_{max} in transmission delay between any two frequencies in the channel band. In the case of linear delay distortion $d_{max} = 2d$, while in the case of quadratic delay distortion $d_{max} = d$, where d is defined in Fig. 11. A third choice might have been the difference in delay d between the midband frequency and the mean sideband frequency $\frac{1}{2}f_{max}$, in which case $d = d/2$ for linear and $d = d/4$ for quadratic delay distortion.

It will be recognized that translation from one basic delay parameter to another can readily be made. Also, the question of whether linear or quadratic delay distortion causes greater transmission impairments will depend significantly on the choice of transmission delay parameters.

6.3 *Double-Sideband AM*

Maximum transmission impairments are shown in Fig. 12 for systems employing $l = 2, 3, 4$ and 5 pulse amplitudes and ideal synchronous detection. With envelope detection the transmission impairments are the same as with synchronous detection, for quadratic delay distortion and for any type of delay distortion with even symmetry about the channel midband (carrier) frequency. However, with envelope detection greater transmission impairments are incurred in the case of linear delay distortion, and for any type of delay distortion with odd symmetry about the channel midband frequency. The difference between envelope and synchronous detection in the presence of linear delay distortion is illustrated in Fig. 12 for $l = 2$ pulse amplitudes.

As noted previously, the maximum transmission impairments indicated in Fig. 12 would be encountered for extremely small error probabilities. For error probabilities in the range normally considered, the maximum impairments given in Fig. 12 would be rather closely approached when the impairments are fairly small, say less than 3 db. However, when the maximum impairments are rather high the actual impairments may be significantly smaller. For example, with a maximum impairment of 10 db, the actual impairment would be expected to be about 1.5 db less for an error probability 10^{-5} and about 2 db less for an error probability 10^{-4}.

6.4 *Vestigial-Sideband AM and Quadrature Double-Sideband AM*

Vestigial-sideband AM and quadrature double-sideband AM with synchronous detection are equivalent methods as regards channel bandwidth requirements and signal-to-noise ratios, in the absence of delay distortion. Both methods may be used in preference to double-sideband AM either (a) to realize a two-fold increase in pulse transmission rate for

a given bandwidth in exchange for a 3 db penalty in signal-to-noise ratio or (b) to secure a two-fold reduction in bandwidth for a given pulse transmission rate, without a penalty in signal-to-noise ratio.

The maximum transmission impairments shown in Fig. 12 are for the same bandwidths as in double-sideband AM with a two-fold increase in the pulse transmission rate. In this case transmission impairments from quadratic delay distortion are no greater than in double-sideband AM, and this applies for any type of delay distortion with even symmetry about the channel midband frequency.

With linear delay distortion, or any delay distortion with odd symmetry about the channel midband frequency, transmission impairments are not identically the same for vestigial-sideband AM and quadrature double-sideband AM. However, the difference is not significant in the case of linear delay distortion, as indicated in Fig. 12. For practical purposes the two methods can be regarded as equivalent for any type of delay distortion actually expected, as regards channel bandwidth requirements and signal-to-noise ratios for a given error probability, assuming ideal synchronous detection.

With linear delay distortion the transmission impairments with the above two methods are significantly greater than for double-sideband AM as indicated by comparison of the curves in Fig. 12 for the two methods for $l = 2$ and 3 pulse amplitudes. This assumes that the pulse transmission rate is twice as great as in double-sideband AM.

When the pulse transmission rate is the same as in double sideband AM but the bandwidth is halved, delay distortion over the channel band is reduced. In this case vestigial-sideband AM or quadrature double-sideband AM affords an advantage over double-sideband AM in the presence of delay distortion with even symmetry about the channel midband frequency, but not necessarily when delay distortion has odd symmetry. With linear delay distortion the ratio d/T is halved, and in this case there is a slight disadvantage compared to double-sideband AM, for $l = 2$ pulse amplitudes. However, with the type of delay distortion ordinarily encountered vestigial-sideband AM and quadrature double-sideband AM would afford some advantage in signal-to-noise ratio over double-sideband AM for equal pulse transmission rates and with ideal synchronous detection.

6.5 *PM with Synchronous Detection*

Two-phase modulation or phase reversal is equivalent to double-sideband AM with equal amplitudes but opposite polarities of the transmitted pulses. The curves in Fig. 12 for double-sideband AM and $l = 2$

pulse amplitudes apply also for two-phase transmission, for the reason that the transmission impairments for a given peak-to-peak difference between pulse amplitudes is the same regardless of polarities.

Two-phase modulation can also be used in conjunction with vestigial-sideband transmission. The curves in Fig. 12 for vestigial-sideband AM and $l = 2$ pulse amplitudes also apply for two-phase vestigial-sideband modulation.

Four-phase modulation is equivalent to bipolar AM on each of two carriers at quadrature with each other. The curves in Fig. 12 for quadrature double sideband AM and $l = 2$ pulse amplitudes also apply for the special case of four-phase modulation.

The maximum transmission impairments with double-sideband two-phase and four-phase modulation and synchronous detection are shown separately in Fig. 12 for comparison with PM with differential phase detection.

6.6 *PM with Differential Phase Detection*

In phase modulation systems differential phase modulation (described in Section 4.1) may be used in place of synchronous detection. Differential phase detection has been implemented in various ways, which in general involve some transmission impairments from channel bandwidth limitations, even with a linear phase characteristic. Such transmission impairments from channel bandwidth limitation is avoided with the implementation assumed herein (Section 4.1), and only the effect of phase distortion is evaluated. Transmission impairments from delay distortion will be greater with this method than with synchronous detection, as illustrated in Fig. 12 for double-sideband two-phase and four-phase quadrature systems and delay distortion. Transmission impairments from linear delay distortion have not been determined for this case.

6.7 *Binary FM*

With optimum systems design, binary FM, or frequency shift keying, requires the same bandwidth for a given pulse transmission rate as binary double-sideband AM. Maximum transmission impairments with these two methods are compared in Fig. 12. It will be noted that with quadratic delay distortion the impairments are smaller with FM than with AM employing either envelope or synchronous detection. In the case of linear delay distortion, the transmission impairments are greater with FM than with synchronous AM, but are somewhat smaller than with AM employing envelope detection.

The transmission impairments given in Fig. 12 for FM apply without a postdetection low-pass filter for noise reduction, and may involve somewhat greater approximations than for the other modulation methods. Approximately the same impairments from phase distortion would be expected with an appropriate low-pass filter.

6.8 *Comparisons of Carrier Modulation Methods*

Signal-to-noise ratios at the detector input for a given error probability and various methods of carrier modulation are ordinarily compared on the premise of ideal amplitude versus frequency characteristics of the channels, and a linear phase characteristic. The curves in Fig. 12 indicate that transmission impairments resulting from phase distortion depend significantly on the carrier modulation method. The optimum method as regards signal-to-noise ratio will thus depend on the type and degree of phase distortion encountered in a particular application. For example, two-phase modulation with synchronous or with differential phase detection may have a slight advantage in signal-to-noise ratio over binary frequency shift keying in the absence of delay distortion. However, the advantage in signal-to-noise ratio would be expected to be with frequency shift keying in application to channels with pronounced quadratic delay distortion or other types of delay distortion with essentially even symmetry about the carrier frequency.

In comparing the performance of various methods of carrier modulation it is necessary to consider other factors than signal-to-noise ratios and channel bandwidth requirements as discussed here. Among them can be mentioned the adverse effects of sudden or gradual level and phase variations and the complexity of instrumentation.

VII. ACKNOWLEDGMENTS

The writer is indebted to A. P. Stamboulis for pointing out some errors in the original equation (50) and for showing the presence of the third term in (56) and to C. F. Pease for numerical evaluation of integrals in the Appendix with the aid of a 704 digital computer.

APPENDIX

Determination of Carrier Pulse Transmission Characteristics

As mentioned in Section 2.2, the in-phase and quadrature components of the carrier pulse transmission characteristics for any carrier frequency

ω_c can be determined from those for any other carrier frequency ω_0, for example the midband frequency of the channel. Basic Fourier integrals are given here for the carrier pulse transmission characteristics for a reference or carrier frequency ω_0. In addition, special integrals are given, applying for a raised cosine pulse spectrum with linear delay distortion, quadratic delay distortion and the type of delay distortion introduced by flat bandpass filters with sharp cutoffs. For these three cases the carrier pulse transmission characteristics have been determined by numerical integration and are tabulated here.

A.1 *General Formulation*

The shape of $R_0(t)$ and $Q_0(t)$ depends on the shape of the transmitted carrier pulse and on the transmission-frequency characteristic of the channel. If the carrier pulse is assumed of sufficiently short duration, the spectrum will be essentially flat over the channel band, so that the shape of the received spectrum is the same as that of the amplitude characteristic of the channel. The functions R_0 and Q_0 are then obtained from expression given elsewhere (Ref. 2, Section 2) in terms of the amplitude characteristic $A(u)$ of the channel, where u is the frequency measured from the carrier frequency ω_0, as indicated in Fig. 13. In the more general case of carrier pulses of any shape and any channel transmission-frequency characteristic, the functions R_0 and Q_0 are obtained by replacing in the above expressions $A(u)$ with the spectrum $S_0(u)$ of the pulse

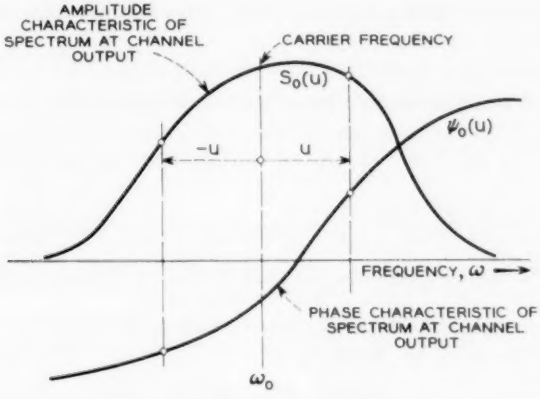


Fig. 13 — Amplitude characteristic $S_0(u)$ and phase characteristic $\Psi_0(u)$ of pulse spectrum at channel output (i.e., detector input) for carrier at frequency ω_0.

envelope at the channel output (detector input). The following expressions are thus obtained in place of (2.10) and (2.11) of Ref. 2:

$$R_0 = R_0^- + R_0^+, \qquad Q_0 = Q_0^- - Q_0^+, \tag{159}$$

$$R_0^- = \frac{1}{\pi}\int_0^{\omega_0} S_0(-u)\cos\left[ut + \Psi_0(-u)\right] du, \tag{160}$$

$$R_0^+ = \frac{1}{\pi}\int_0^{\infty} S_0(u)\cos\left[ut - \Psi_0(u)\right] du, \tag{161}$$

$$Q_0^- = \frac{1}{\pi}\int_0^{\omega_0} S_0(-u)\sin\left[ut + \Psi_0(-u)\right] du, \tag{162}$$

$$Q_0^+ = \frac{1}{\pi}\int_0^{\infty} S_0(u)\sin\left[ut - \Psi_0(u)\right] du. \tag{163}$$

The various quantities in the above expressions are as shown in Fig. 13. It will be recognized that the upper limit ω_0 in (160) and (162) can for practical purposes be replaced by ∞, since $S_0(-\omega_0) \cong 0$.

A.2 *Even Symmetry Spectrum and Delay Distortion*

Let the spectrum at the detector input have even symmetry about ω_0 and the phase distortion odd symmetry, in which case

$$S_0(-u) = S_0(u), \tag{164}$$

$$\Psi_0(-u) = -\Psi_0(u). \tag{165}$$

Delay distortion will then have even symmetry about ω_0, i.e., $\Psi_0'(-u) = \Psi_0'(u)$.

With (164) and (165) in (159) through (163), the following relations are obtained when the upper limit ω_0 is replaced by ∞:

$$R_0(t) = \frac{2}{\pi}\int_0^{\infty} S_0(u)\cos\left[ut - \Psi_0(u)\right] du, \tag{166}$$

$$Q_0(t) = 0. \tag{167}$$

A.3 *Even Symmetry Spectrum and Odd Symmetry Delay Distortion*

When the phase characteristic has a component with even symmetry about the frequency ω_0, so that

$$\Psi_0(-u) = \Psi_0(u) \tag{168}$$

the corresponding delay distortion will have odd symmetry.

With (164) and (168) in (159) through (163), the following relations are obtained:

$$R_0(-t) = R_0(t) = \frac{2}{\pi}\int_0^\infty S_0(u)\cos ut\cos\Psi_0(u)\,du, \quad (169)$$

$$Q_0(-t) = Q_0(t) = \frac{2}{\pi}\int_0^\infty S_0(u)\cos ut\sin\Psi_0(u)\,du. \quad (170)$$

A.4 *Raised Cosine Pulse Spectrum*

For reasons discussed elsewhere (Ref. 2, Section 5) it is desirable in pulse systems to employ raised cosine pulse spectra, as shown in Fig. 1 and given by

$$S_0(-u) = S_0(u) = \frac{T}{2}\cos^2\frac{\pi u}{4\bar{\omega}}, \quad (171)$$

where $\bar{\omega}$ is the mean frequency from midband.

The corresponding carrier pulse transmission characteristic obtained from (159) through (163) with $\Psi_0(u) = 0$ is

$$\bar{P}_0 = R_0(t) = \frac{\sin 2\bar{\omega}t}{2\bar{\omega}t[1 - (2\bar{\omega}t/\pi)^2]}. \quad (172)$$

Pulses can in this case be transmitted without intersymbol interference at intervals T such that

$$\bar{\omega}T = \pi. \quad (173)$$

A.5 *Quadratic Delay Distortion*

It will be assumed that the phase characteristic contains a linear component, which can be disregarded, and a distortion component given by

$$\Psi_0(u) = cu^3, \quad (174)$$

where c is a constant. The corresponding delay distortion is then quadratic or parabolic, as given by

$$\Psi_0'(u) = 3cu^2. \quad (175)$$

In this case $Q_0(t) = 0$ in accordance with (167), while (175) in (166) gives

$$R_0(t) = \frac{4}{\pi}\int_0^{\pi/2}\cos^2 x\cos(ax - bx^3)\,dx, \quad (176)$$

TABLE XIX — FUNCTIONS $R_0(t/T)$ AND $R_1(t/T)$ FOR RAISED COSINE PULSE SPECTRUM AND QUADRATIC DELAY DISTORTION

t/T	$d/T = 0$		$d/T = 1$		$d/T = 2$		$d/T = 3$		$d/T = 4$	
	R_0	$-R_1$	R_0	$-R_1$	R_0	$-R_1$	R_0	$-R_1$	R_0	$-R_1$
−3.00	0	−0.0024	−0.0006	0.0005	0.0025	0.0028	0	−0.0003	0	0
−2.75	0.0020	−0.0006	−0.0003	−0.0011	−0.0023	0.0048	−0.0002	0	0	−0.0003
−2.50	0	0.0042	0.0009	−0.0007	−0.0030	−0.0004	0	0.0002	0.0005	−0.0005
−2.25	−0.0037	0.0013	0.0004	0.0017	0.0005	−0.0009	0.0002	−0.0007	0.0011	−0.0009
−2.00	0	−0.0083	−0.0013	0.0009	0.0011	0.0003	0.0017	−0.0024	0.0028	−0.0029
−1.75	0.0081	−0.0037	0	−0.0038	0.0020	−0.0023	0.0055	−0.0055	0.0079	−0.0079
−1.50	0	0.0208	0.0053	−0.0062	0.0081	−0.0114	0.0145	−0.0139	0.0202	−0.0178
−1.25	−0.0243	0.0164	0.0137	−0.0138	0.0280	−0.0306	0.0374	−0.0344	0.0464	−0.0367
−1.00	0	−0.0833	0.0467	−0.0627	0.0756	−0.0689	0.0891	−0.0721	0.0986	−0.0703
−0.75	0.1698	−0.2603	0.1699	−0.1774	0.1772	−0.1404	0.1879	−0.1282	0.1923	−0.1192
−0.50	0.5000	−0.3750	0.4064	−0.3036	0.3645	−0.2342	0.3492	−0.1943	0.3389	−0.1731
−0.25	0.8488	−0.2829	0.7297	−0.3126	0.6303	−0.2813	0.5692	−0.2373	0.5320	−0.2070
0	1.00	0	0.9633	−0.1228	0.8795	−0.1884	0.7956	−0.1963	0.7336	−0.1830
0.25	0.8488	0.2829	0.9357	0.1809	0.9540	0.0580	0.9182	−0.0276	0.8666	−0.0661
0.50	0.5000	0.3750	0.6338	0.3900	0.7557	0.3255	0.8223	0.2227	0.8364	0.1361
0.75	0.1698	0.2603	0.2366	0.3638	0.3610	0.4237	0.4952	0.4056	0.5936	0.3386
1.00	0	0.0833	−0.0341	0.1634	−0.0098	0.2838	0.0827	0.3792	0.2045	0.4073
1.25	−0.0243	−0.0164	−0.0954	−0.0236	−0.1700	0.0367	−0.1951	0.1547	−0.1501	0.2692
1.50	0	−0.0208	−0.0341	−0.0752	−0.1161	−0.1165	−0.2200	−0.0898	−0.2909	0.0065
1.75	0.0081	0.0037	0.0205	−0.0260	0.0038	−0.0985	−0.0714	−0.1729	−0.1869	−0.1872
2.00	0	0.0083	0.0196	0.0205	0.0543	−0.0016	0.0655	−0.0809	0.0142	−0.1802
2.25	−0.0037	−0.0013	−0.0031	0.0176	0.0242	0.0477	0.0799	0.0436	0.1237	−0.0283
2.50	0	−0.0042	−0.0091	−0.0046	−0.0151	0.0225	0.0120	0.0727	0.0814	0.0945
2.75	0.0020	0.0006	−0.0002	−0.0093	−0.0170	−0.0150	−0.0357	0.0158	−0.0185	0.0834
3.00	0	−0.0024	0.0044	0.0007	0.0020	−0.0167	−0.0231	−0.0327	−0.0584	−0.0583
3.25	−0.0012	−0.0028	0.0061	0.0050	0.0091	0.0028	0.0093	−0.0234	−0.0216	−0.0543
3.50	0	−0.0115	−0.0024	0.0002	0.0011	0.0097	0.0159	0.0090	0.0229	−0.0249
3.75	0.0007	0.0002	−0.0005	−0.0029	−0.0047	0.0007	0.0003	0.0164	0.0229	0.0208
4.00	0	0.0010	0.0014	−0.0004	−0.0014	−0.0053	−0.0087	0	−0.0037	0.0241
4.25	−0.0005	0	0.0004	−0.0017	0.0024	−0.0013	−0.0025	−0.0095	−0.0147	−0.0030
4.50	0	−0.0007	−0.0009	0.0003	0.0011	0.0030	0.0044	−0.0025	−0.0030	−0.0157
4.75	0.0004	0	−0.0003	−0.0011	−0.0014	0.0012	0.0024	0.0051	0.0078	−0.0033
5.00	0	0.0005	0.0006	−0.0003	−0.0008	−0.0008	−0.0022	0.0025	0.0040	0.0086

where

$$a = 4\,\frac{t}{T}, \qquad b = \frac{16}{3\pi^2}\,\frac{d}{T}. \tag{177}$$

The ratio t/T is the time measured in pulse intervals and the ratio d/T the maximum delay distortion measured in pulse intervals, with d defined as in Fig. 2 or Fig. 11.

In certain cases, as in connection with pulse transmission by frequency modulation, the time derivative of $R_0(t)$ is involved. This derivative is given by

$$dR_0/dt = \frac{4}{T}\,R_1(t), \tag{178}$$

where

$$R_1(t) = dR/da$$

and is given by

$$R_1(t) = -\frac{4}{\pi}\int_0^{\pi/2} x\cos^2 x \sin\,(ax - bx^3)\,dx. \tag{179}$$

The functions $R_0(t)$ and $R_1(t)$ obtained by numerical integration of (176) and (179) are given in Table XIX. The function $R_0(t/T)$ is shown in Fig. 2.

A.6 *Linear Delay Distortion*

It will be assumed that the phase distortion component is given by

$$\Psi_0(u) = cu^2, \tag{180}$$

which corresponds to a linear delay distortion given by

$$\Psi_0'(u) = 2cu. \tag{181}$$

In this case expressions (169) and (170) give

$$R_0(-t) = R_0(t) = \frac{4}{\pi}\int_0^{\pi/2} \cos^2 x \cos ax \cos bx^2\,dx, \tag{182}$$

$$Q_0(-t) = Q_0(t) = \frac{4}{\pi}\int_0^{\pi/2} \cos^2 x \cos ax \sin bx^2\,dx, \tag{183}$$

where

$$a = 4\,\frac{t}{T}, \qquad b = \frac{4}{\pi}\,\frac{d}{T},$$

in which the delay d is defined as in Fig. 4 or Fig. 11.

TABLE XX — FUNCTIONS $R_0(t/T)$ AND $Q_0(t/T)$ FOR RAISED COSINE PULSE SPECTRUM AND LINEAR DELAY DISTORTION

t/T	$d/T = 0$		$d/T = 0.5$		$d/T = 1$		$d/T = 2$		$d/T = 3$	
	R_0	Q_0	R_0	Q_0	R_0	Q_0	R_0	Q_0	R_0	Q_0
0	1.00	0	0.9516	0.1941	0.8309	0.3306	0.5786	0.3895	0.4537	0.3471
±0.25	0.8488	0	0.8235	0.1277	0.7567	0.2293	0.5871	0.3180	0.4668	0.3138
±0.50	0.5000	0	0.5181	−0.0105	0.5563	0.0096	0.5733	0.1304	0.4921	0.2106
±0.75	0.1698	0	0.2064	−0.1029	0.2968	−0.1592	0.4674	−0.0934	0.4878	0.0423
±1.00	0	0	0.0180	−0.0956	0.0726	−0.1792	0.2596	−0.2439	0.4026	−0.1489
±1.25	−0.0243	0	−0.0331	−0.0336	−0.0447	−0.0917	0.0327	−0.2488	0.2215	−0.2823
±1.50	0	0	−0.0155	0.0101	−0.0538	−0.0030	−0.1035	−0.1334	0.0029	−0.2858
±1.75	0.0081	0	0.0039	0.0122	−0.0147	0.0257	−0.1087	−0.0014	−0.1472	−0.1620
±2.00	0	0	0.0048	−0.0014	0.0112	0.0107	−0.0383	0.0577	−0.1614	−0.0004
±2.25	−0.0037	0	−0.0007	−0.0049	0.0085	−0.0051	0.0216	0.0380	−0.0709	0.0911
±2.50	0	0	−0.0018	0.0003	−0.0023	−0.0048	0.0269	−0.0025	0.0255	0.0767
±2.75	0.0020	0	0.0002	0.0024	−0.0039	0.0021	0.0043	−0.0164	0.0543	0.0109
±3.00	0	0	0.0008	0	0.0006	0.0024	−0.0094	−0.0052	0.0229	−0.0300
±3.25	−0.0012	0							−0.0126	−0.0221
±3.50	0	0							−0.0169	−0.0031
±3.75	0.0007	0							−0.0139	0.0116

The values of R_0 and Q_0 obtained by numerical integration of (182) and (183) are given in Table XX.

It will be noted that $Q_0(0) \neq 0$. From the standpoint of analysis, it may be convenient to modify the phase such that $Q_0(0) = 0$. The modified values are given by

$$\begin{aligned} R_{00}(t) &= [R_0^2(t) + Q_0^2(t)]^{\frac{1}{2}} \cos [\Psi_0(t) - \Psi_0(0)] \\ &= k_1 R_0(t) + k_2 Q_0(t), \end{aligned} \tag{184}$$

$$\begin{aligned} Q_{00}(t) &= [R_0^2(t) + Q_0^2(t)]^{\frac{1}{2}} \sin [\Psi_0(t) - \Psi_0(0)] \\ &= k_1 Q_0(t) - k_2 R_0(t), \end{aligned} \tag{185}$$

where

$$\begin{aligned} k_1 &= \frac{R_0(0)}{[R_0^2(0) + Q_0^2(0)]^{\frac{1}{2}}}, \\ k_2 &= \frac{Q_0(0)}{[R_0^2(0) + Q_0^2(0)]^{\frac{1}{2}}}. \end{aligned} \tag{186}$$

The modified values are given in Table XXI. The functions $R_{00}(t/T)$ and $Q_{00}(t/T)$ are shown in Fig. 4.

The time derivatives of $R_0(t)$ and $Q_0(t)$ are of interest in connection with frequency modulation and given by

$$dR_0/dt = \frac{4}{T} dR_0/da = \frac{4}{T} R_1(t), \tag{187}$$

$$dQ_0/dt = \frac{4}{T} dQ_0/da = \frac{4}{T} Q_1(t), \tag{188}$$

where

$$R_1(t) = -\frac{4}{\pi} \int_0^{\pi/2} x \cos^2 x \sin ax \cos bx^2 \, dx, \tag{189}$$

$$Q_1(t) = -\frac{4}{\pi} \int_0^{\pi/2} x \cos^2 x \sin ax \sin bx^2 \, dx. \tag{190}$$

The functions R_1 and Q_1 obtained by numerical integration are given in Table XXII.

The following functions occur in connection with binary FM:

$$\frac{1}{\bar{\omega}} \frac{dR_0(t)}{dt} = \frac{4}{\bar{\omega}T} R_1(t) = \frac{4}{\pi} R_1(t), \tag{191}$$

$$\frac{1}{\bar{\omega}} \frac{dQ_0(t)}{dt} = \frac{1}{\bar{\omega}T} Q_1(t) = \frac{4}{\pi} Q_1(t). \tag{192}$$

These functions are given in Table XVI for integral values of $n = t/T$.

TABLE XXI — FUNCTIONS $R_{00}(t/T)$ AND $Q_{00}(t/T)$ FOR RAISED COSINE PULSE SPECTRUM AND LINEAR DELAY DISTORTION

t/T	$d/T = 0$		$d/T = 0.5$		$d/T = 1$		$d/T = 2$		$d/T = 3$	
	R_{00}	Q_{00}	R_{00}	Q_{00}	R_{00}	Q_{00}	R_{00}	Q_{00}	R_{00}	Q_{00}
0	1.00		0.9712	0	0.8943	0	0.6975	0	0.5712	0
±0.25	0.8488	0	0.8324	−0.0395	0.7879	−0.0667	0.6646	−0.0641	0.5614	−0.0344
±0.50	0.5000	0	0.5056	−0.1138	0.5204	−0.1967	0.5484	−0.2120	0.5188	−0.1317
±0.75	0.1698	0	0.1817	−0.1421	0.2169	−0.2577	0.3356	−0.3385	0.4131	−0.2628
±1.00	0	0	−0.0015	−0.0973	0.0012	−0.1933	0.0792	−0.3473	0.2293	−0.3629
±1.25	−0.0243	0	−0.0392	−0.0263	−0.0754	−0.0687	−0.1118	−0.2247	0.0044	−0.3588
±1.50	0	0	−0.0132	0.0130	−0.0511	0.0171	−0.1604	−0.0529	−0.1714	−0.2288
±1.75	0.0081	0	0.0063	0.0111	−0.0042	0.0293	−0.0910	0.0595	−0.2153	−0.0392
±2.00	0	0	0.0044	−0.0023	0.0144	0.0058	0.0005	0.0693	−0.1284	0.0978
±2.25	−0.0037	0	−0.0017	−0.0047	0.0060	−0.0079	0.0391	0.0195	−0.0010	0.1154
±2.50	0	0	−0.0017	0.0007	−0.0039	−0.0036	0.0209	−0.0171	0.0669	0.0454
±2.75	0.0020	0	0.0007	−0.0023	−0.0032	0.0026	0.0056	−0.0160	0.0498	−0.0243
±3.00	0	0	0.0008	−0.0002	0.0014	0.0020	−0.0107	0.0009	−0.0000	−0.0377
±3.25	−0.0012	0							−0.0234	−0.0100
±3.50	0								−0.0115	0.0127
±3.75	0.0007								−0.0040	0.0177

TABLE XXII — FUNCTIONS $R_1(t/T)$ AND $Q_1(t/T)$ FOR RAISED COSINE PULSE SPECTRUM AND LINEAR DELAY DISTORTION

$$R_1(-t/T) = -R_1(t/T); \qquad Q_1(-t/T) = -Q_1(t/T)$$

t/T	$d/T = 0$		$d/T = 0.5$		$d/T = 1.0$		$d/T = 2.0$		$d/T = 3.0$	
	R_1	Q_1	R_1	Q_1	R_1	Q_1	R_1	Q_1	R_1	Q_1
0	0	0	0	0	0	0	0	0	0	0
0.25	−0.2829	0	−0.2418	−0.1214	−0.1448	−0.1869	+0.0095	−0.1385	+0.0239	−0.0674
0.50	−0.3750	0	−0.3379	−0.1329	−0.2446	−0.2212	−0.0505	−0.2231	+0.0194	−0.1389
0.75	−0.2603	0	−0.2626	−0.0423	−0.2578	−0.0987	−0.1637	−0.2051	−0.0373	−0.1907
1.00	−0.0833	0	−0.1125	+0.0490	−0.1777	+0.0502	−0.2369	−0.0827	−0.1363	−0.1774
1.25	+0.0164	0	−0.2872	+0.0624	−0.0568	+0.1050	−0.1967	+0.0683	−0.2158	−0.0761
1.50	+0.0208	0	+0.0261	+0.0212	+0.0274	+0.0617	−0.0686	+0.1434	−0.2020	+0.0689
1.75	−0.0037	0	+0.0097	−0.0118	+0.0396	−0.0004	+0.0477	+0.1045	−0.0859	+0.1625
2.00	−0.0083	0	−0.0051	−0.0105	+0.0010	−0.0214	+0.0770	+0.0131	+0.0517	+0.1408
2.25	+0.0013	0	−0.0040	+0.0029	−0.0112	−0.0075	+0.0330	−0.0430	+0.1110	+0.0359
2.50	+0.0042	0	+0.0014	+0.0051	−0.0073	+0.0058	−0.0150	−0.0311	+0.0684	−0.0541
2.75	−0.0006	0	+0.0018	−0.0010	+0.0032	+0.0042	−0.0227	+0.0025	−0.0089	−0.0633
3.00	−0.0024	0	0.0005	−0.0028	+0.0040	−0.0018	−0.0033	+0.0148	−0.0430	−0.0146
3.25									−0.0216	−0.0240
3.50									+0.0103	+0.0201
3.75									+0.0157	−0.0028
4.00									+0.0011	−0.0112

A.7 *Delay Distortion from Flat Bandpass Filters*

Let a bandpass filter have an amplitude characteristic A_0 between $-\omega_c + \omega_0$ and $\omega_0 + \omega_c$ and A_1 outside this band. When the bandwidth $2\omega_c$ is small in relation to the midband frequency ω_0, the phase characteristic is closely approximated by

$$\psi_0(u) = \frac{B}{\pi} \log_e \frac{1 + u/\omega_c}{1 - u/\omega_c}, \tag{193}$$

where

$$B = \log_e (A_0/A_1). \tag{194}$$

The corresponding envelope delay distortion is $D(u) = d\psi_0(u)/du$ and delay distortion relative to the midband frequency becomes

$$D_0(u) = D(u) - D(0) = \frac{2B}{\pi\omega_c} \frac{(u/\omega_c)^2}{1 - (u/\omega_c)^2} \tag{195}$$

$$= \frac{2B}{\pi\omega_c}\left[\left(\frac{u}{\omega_c}\right)^2 + \left(\frac{u}{\omega_c}\right)^4 + \left(\frac{u}{\omega_c}\right)^6 + \cdots\right]. \tag{196}$$

It will be noted that the first term in (196) represents quadratic delay distortion, which is approximated for $u/\omega_c \ll 1$.

Let the pulse spectrum at the detector input have a raised cosine shape, as given by (171), in which case the maximum radian frequency to each side of midband is $2\bar{\omega}$. With a phase characteristic as given by (193), the carrier pulse transmission characteristic is in this case obtained with (171) and (193) in (166) and becomes

$$R_0(-t) = R_0(t) = \frac{4}{\pi}\int_0^{\pi/2} \cos^2 x \cos\left[ax - \psi_0(x)\right] dx, \tag{197}$$

where

$$a = 4t/T,$$

$$\psi_0(x) = \frac{B}{\pi} \log_e\left(\frac{k + \frac{2}{\pi}x}{k - \frac{2}{\pi}x}\right), \tag{198}$$

$$k = \frac{\omega_c}{2\bar{\omega}} = \frac{W_2}{W_1}, \tag{199}$$

in which W_1 is the bandwidth of the raised cosine spectrum and W_2 that of the flat filter, as indicated in Fig. 14.

In Table XXIII are given the values of $R_0(t/T)$ obtained by numeri-

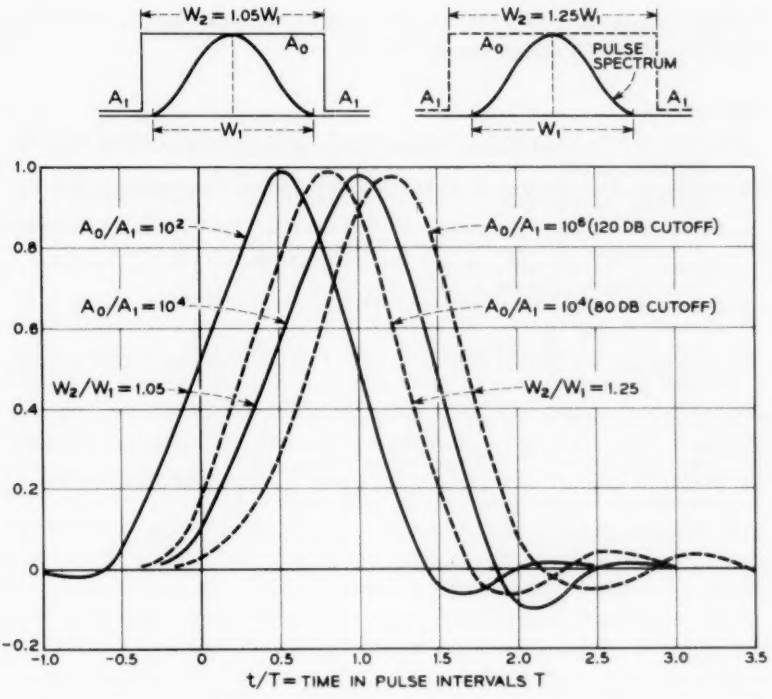


Fig. 14 — Carrier pulse transmission characteristics for raised cosine pulse spectrum and phase distortion resulting from flat filters with sharp cutoffs.

TABLE XXIII — FUNCTION $R_0(t/T)$ FOR RAISED COSINE SPECTRUM AND PHASE DISTORTION RESULTING FROM FLAT FILTERS WITH SHARP CUTOFFS

$k = W_2/W_1$		1.05		1.25	
A_0/A_1		10^2	10^4	10^4	10^6
A_0/A_1, in db		40	80	80	120
t/T	−1.0	−0.002	∼0	0.001	∼0
	−0.5	0.048	0.001	−0.001	∼0
	0	0.525	0.092	0.168	0.018
	0.5	0.994	0.554	0.758	0.285
	1.0	0.481	0.979	0.903	0.888
	1.5	−0.051	0.464	0.192	0.789
	2.0	0.004	−0.110	−0.060	0.059
	2.5	0.003	0.005	0.020	−0.059
	3.0	−0.003	0.001	−0.008	0.028

cal integration of (197) for certain cases as indicated in the table. The functions $R_0(t/T)$ are shown in Fig. 14.

REFERENCES

1. Sunde, E. D., Ideal Binary Pulse Transmission by AM and FM, B.S.T.J., **38,** 1959, p. 1357.
2. Sunde, E. D., Theoretical Fundamentals of Pulse Transmission, B.S.T.J., **33,** 1954, pp. 721; 987.
3. Gibby, R. A., An Evaluation of AM Data System Performance by Computer Simulation, B.S.T.J., **39,** 1960, p. 675.
4. Fowler, A. D. and Gibby, R. A., Assessment of Effects of Delay Distortion in Data Systems, Comm. & Elect., no. 40, 1959, p. 918.
5. Costas, J. P., Synchronous Communications, Proc. I.R.E., **44,** 1956, p. 1713.
6. Rieke, J. W. and Graham, R. S., The L-3 Coaxial System — Television Terminals, B.S.T.J., **32,** 1953, p. 915.
7. Darlington, S., Realization of Constant Phase Difference, B.S.T.J., **29,** 1950, p. 94.
8. Cahn, C. R., Performance of Digital Phase Modulation Communication Systems, Proc. I.R.E. Trans., **CS-7,** 1959, p. 3.
9. Cahn, C. R., Combined Digital Phase and Amplitude Modulation Communication Systems, I.R.E. Trans., **CS-8,** 1960, p. 150.
10. Mosier, R. R. and Clabaugh, R. G., Kineplex — A Bandwidth Efficient Binary System, Comm. & Elect., no. 34, 1958, p. 723.

Further Results on the Detectability of Known Signals in Gaussian Noise

By H. C. MARTEL and M. V. MATHEWS

(Manuscript received September 12, 1960)

The detection of a completely known signal which may or may not be present in a finite sample of gaussian noise is considered from two points of view. The first examines the performance of a maximum likelihood detector operating on a finite set of discrete measurements of the stimulus as the set becomes large. The stimulus is either signal plus noise or noise alone. Examples are presented for signals in bandlimited noise, using as measurements either equispaced amplitude samples or derivatives at one instant in time. For both, the detectability grows without bound as the number of measurements is increased. The second point of view bases detection on a continuous measurement (linear integral operator) which maximizes the detectability. Solutions have been obtained when the noise has a rational power spectral density. The detector utilizes a cross-correlation between stimulus and signal which is well known and a mechanism, designated extrapolation detection, which involves evaluation of derivatives of the stimulus. The contribution of the derivative measurements to the detectability is examined as the noise approaches bandlimited noise and is found in many cases to grow without bound.

I. INTRODUCTION

The problem under consideration here is the detection of a completely known signal which may or may not be present in a finite sample of gaussian noise. That is, we imagine a situation similar to Fig. 1 in which a stimulus is made up of either signal plus noise or noise alone and we ask, given T seconds of this stimulus, how accurately can we decide whether or not the signal is present. The noise is thought of as having been produced by a stochastic process and thus the question is really one of statistical hypothesis testing.

This particular problem has been treated rather extensively,[1-7] and certain questions, even controversies, have arisen. These concern what

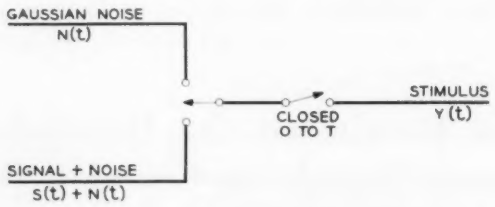


Fig. 1 — Diagram of problem under consideration.

constitutes a proper description for the stimulus, under what circumstances can the stimulus be characterized by a finite number of samples, and under what conditions is perfect detectability obtained, i.e., when is it always possible to detect the presence or absence of the signal. Peterson, Birdsall and Fox[6] have described the stimulus as being *Fourier series bandlimited* and by so doing have obtained quite different results from the other authors, who for the most part consider stationary gaussian noise. In many cases, finite-duration stimuli have been characterized by a finite number of samples usually chosen so they are independent, and maximum likelihood detectors operating on these samples have been developed. This has led to the equivalent of a correlation detection process in which the test statistic is the integral of the product of the stimulus and a function derived from the signal. Such detectors always produce finite detectability. On the other hand, Slepian[7] has pointed out by an argument involving analytic continuation that many signals can be perfectly separated from noise provided the noise is considered to have a bandlimited spectrum. Clearly some mechanism in addition to correlation detection is inherent in Slepian's result, and indeed he points out one such detector.

The results of Peterson, Birdsall and Fox have been used extensively for comparison with the performance achieved by humans and other animals, and questions as to the validity of such comparisons originally motivated this investigation. However, it seems very doubtful if the mechanisms which will be developed can have anything to do with perception. In addition, we have chosen to work with stationary gaussian noise rather than Fourier series bandlimited noise, the former being a much more satisfactory characterization of real noise.

Two different attempts to better understand the questions cited above have been undertaken. The first examines the performance of a maximum likelihood detector operating on a finite set of discrete measurements of the stimulus as the set becomes very large. The results show cases where the detectability grows without bound. Thus, the charac-

terization of the stimulus by a finite set of measurements is incomplete. However, in some cases, a law of diminishing returns operates so that the rate of increase in detectability slows as the number of samples is increased.

The second study bases detection on a continuous measurement (linear integral operator), which is the solution of an optimizing integral equation. The test statistic so obtained has two parts, one similar to correlation detection, the other based on measurements of the derivatives of the stimulus. The contribution of this latter term is usually the smaller of the two, but, where the noise spectrum approaches a bandlimited form, it may grow without bound. In addition, it may be important if the stimulus is very short.

Both maximum likelihood detection with a finite number of samples and the integral equation for the continuous statistic have been previously presented. The new contributions arise from the more complete solutions which have been obtained. The most significant result is undoubtedly the solution of the integral equation in closed form so that its characteristics and particularly its asymptotic properties for many-pole noise can be seen. The derivative detector, which will be termed *extrapolation detection*, was apparent from this solution.

II. DETECTION WITH A FINITE NUMBER OF SAMPLES

In this section we will derive the maximum likelihood detector for detecting a known signal in gaussian noise from a finite number of samples of the stimulus and apply this detector to two specific problems involving bandlimited noise. Each sample results from some linear operation on the stimulus and the samples need not be independent. The derivation of the detection equation differs only slightly from previously published work,[5] and is included to lead clearly into the specific problems, which are the principal new results. In the problems the behavior of the detector is studied as the number of samples becomes large, first when the samples consist simply of amplitude measurements of the stimulus and second when the samples are a set of derivatives at one point in time.

2.1 *Maximum Likelihood Detector*

The stimulus

$$Y(t) = \begin{cases} N(t) \\ N(t) + S(t) \end{cases} \qquad 0 \leqq t \leqq T, \tag{1}$$

is either a gaussian noise $N(t)$ or that noise plus a known signal $S(t)$ and is observed for the interval $0 \leqq t \leqq T$. The n samples,

$$Y_1 , Y_2 , \cdots , Y_n ,$$

on which the detection is made are obtained by n linear operations $L_1 , L_2 , \cdots , L_n$ on the stimulus

$$Y_i = L_i[Y(t)] \qquad i = 1, \cdots , n.$$

Because of their linearity,

$$L_i[N(t) + S(t)] = L_i[N(t)] + L_i[S(t)] \equiv N_i + S_i ,$$

and N_i will be gaussian random variables which may be completely characterized by their matrix β of correlation coefficients,

$$\beta_{ij} = E\langle N_i N_j \rangle ,$$

and by their means which for simplicity will be assumed to be zero,

$$E\langle N_i \rangle = 0.$$

The density function of the Y_i samples when the stimulus is noise alone may then be written

$$f_N(y_1 , \cdots , y_n) = (2\pi)^{-n/2} \mid \beta \mid^{-\frac{1}{2}} \exp \{ -\tfrac{1}{2} \sum_{i,j} \beta_{ij}^{-1} y_i y_j \},$$

where $y_1 , \cdots , y_n$ are the dummy arguments of the density function corresponding to $Y_1 , \cdots , Y_n$ and $\mid \beta \mid$ is the determinant of β, with all sums going over the range 1 to n unless especially indicated otherwise. The density function of Y_i for signal plus noise is simply

$$f_{SN}(y_1 , \cdots , y_n) = f_N(y_1 - S_1 , \cdots , y_n - S_n)$$

because the signal is additive. Thus the likelihood ratio $L(y_1 , \cdots , y_n)$ is

$$L(y_1 , \cdots , y_n) = \frac{f_{SN}(y_1 , \cdots , y_n)}{f_N(y_1 , \cdots , y_n)} ,$$

which when evaluated for these density functions becomes

$$L(y_1 , \cdots , y_n) = \exp \{ -\tfrac{1}{2} \sum_{i,j} \beta_{ij}^{-1} S_i S_j \} \exp \{ \sum_{i,j} \beta_{ij}^{-1} S_i y_j \}.$$

A maximum likelihood detector says that signal is present if test statistic $L(Y_1 , \cdots , Y_n)$ is greater than some threshold α and will maximize the conditional probability of detecting a signal when it is present for a given conditional probability of indicating signal for noise alone.

However, L is a monotonic function of the statistic φ,

$$\varphi = \sum_{i,j} \beta_{ij}^{-1} S_i Y_j , \tag{2}$$

and consequently an equally good test is $\varphi > \alpha_c$, where α_c is an equivalent threshold. φ may be characterized by two density functions, one if the stimulus is noise alone, the other for signal plus noise. For noise alone, φ_N (the subscript "N" designates noise alone, "SN" signal plus noise) is gaussian with zero mean and variance

$$E\langle \varphi_N^2 \rangle = \sum_{i,j} \beta_{ij}^{-1} S_i S_j .$$

For signal plus noise φ_{SN} is also gaussian with the same variance but with mean

$$E\langle \varphi_{SN} \rangle = \sum_{i,j} \beta_{ij}^{-1} S_i S_j .$$

The density functions for φ are pictured on Fig. 2. The effectiveness of this detector as indicated by the signal detection probability at a given false alarm rate can be characterized by a single number d, which is the ratio of the squared mean of the signal plus noise distribution to the variance of either distribution. The larger d is, the more completely separated are the distributions on Fig. 2 and the higher will be the detection probability. This number d is then

$$d = \sum_{i,j} \beta_{ij}^{-1} S_i S_j . \tag{3}$$

An alternate form for the statistic φ from that given in (2) is

$$\varphi = \sum_j Z_j Y_j , \tag{4}$$

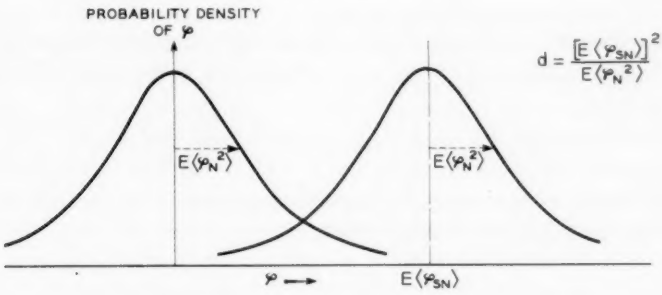


Fig. 2 — Two density functions characterizing statistic φ.

where the Z_j's are solutions to the equations

$$\sum_i \beta_{ij} Z_i = S_j \qquad j = 1, \cdots, n \tag{5}$$

and d may be expressed

$$d = \sum_j Z_j S_j \,. \tag{6}$$

This form is usually preferable for computations since it involves the solution of n linear equations rather than the inversion of an $n \times n$ matrix. In addition this form more closely resembles the integrals which will appear when continuous statistics are considered.

To summarize, a statistic φ which operates on a set of n correlated samples and which is equivalent to a maximum likelihood statistic has been developed. Signal is indicated if φ is greater than some threshold. φ is formed as a linear sum of the samples, it has a gaussian distribution, and it has the same variance for both noise alone and signal plus noise cases. The performance of the detector may be characterized by a single number $d = [E\langle\varphi_{SN}\rangle]^2 / E\langle\varphi_N{}^2\rangle$, the larger the d, the better the performance.

2.2 *Detection of Sinusoid in Bandlimited Noise with Time Samples*

The argument presented by Slepian[7] indicates that theoretically, because of the analytic nature of the noise, a sinusoid can always be detected in spectral bandlimited noise. However, this result says nothing about how fast the detectability increases with the complexity of the detector. In this section an example is examined in which the stimulus is time sampled with n samples equally spaced over the interval $0 \leqq t \leqq T$ and detectability is computed as a function of n. In addition to the general behavior of this function, it is of special interest to note whether any peculiarities occur at $n = 2WT$ (the Nyquist rate), W being the noise bandwidth, since this is the maximum number of independent samples which may be formed. The correlation function of the noise is

$$R(\tau) = E\langle N(t)N(t+\tau)\rangle = \frac{\sin 2\pi W\tau}{2\pi W\tau},$$

where the noise has unit mean square amplitude so the matrix of correlation coefficients β_{ij} can be written

$$\beta_{ij} = \frac{\sin \pi \dfrac{n_q}{n}(i-j)}{\pi \dfrac{n_q}{n}(i-j)}$$

with

$$n_q = 2WT.$$

Unfortunately, no analytic way for either inverting this matrix or solving (5) is known, hence the detectability was computed numerically. This computation was carried out on an IBM 704 machine for a signal with frequency centered in the noise band

$$S_i = A \sin \frac{\pi}{2}\left(\frac{n_q}{n} i - \frac{n_q}{2n}\right),$$

A being the amplitude and $\pi n_q/4n$ being an arbitrary phase chosen for computational convenience. The normalized results of a solution of (5) and (6) are presented on Fig. 3, where d/A^2 is given as a function of the number of samples n/n_q and of the stimulus duration in terms of the number of independent samples n_q. The curves exhibit a knee, not at $n = n_q$ but for n a bit larger than n_q. Detectability continues to increase but the rate of increase becomes imperceptible. The curves are all carried out to a matrix of size 128 × 128, which is the limit of the capacity of the computer program. Double precision arithmetic and a sufficient error analysis were used to insure the accuracy of the results. The increase in detectability beyond $n = n_q$ is essentially equivalent to that which would be obtained by increasing T to $T + 2/W$ and sampling at the Nyquist rate. Heuristically we can say that, by adding extra points inside the interval, it is quite easy to predict $N(t)$ two independent sample times beyond each end of the interval, but very hard to predict further. In an unpublished proof Slepian has shown that the quadratic form for d given by (3) does become infinite for bandlimited noise as n becomes infinite. However, the present example indicates it increases at an exceedingly slow rate. Clearly a statistic which improves more rapidly is desirable, and such is evaluated in the next section.

2.3 *Detection of a Constant in Bandlimited Noise Using Derivatives*

The solution for the optimum integral operator detector carried out in the next section produced a statistic involving derivatives of the stimulus. This result suggests trying derivatives for bandlimited noise, particularly since all derivatives of a bandlimited stimulus exist. Consequently, the detectability achieved by n samples, which are the stimulus and its $n - 1$ derivatives evaluated at one point in time, is studied. This quantity, as will be seen, has the pleasant characteristics of being analytically rather than only numerically determinable and of increasing uniformly with n rather than exhibiting the knee curves of the time samples. A curious property is that the duration of the stimulus is no

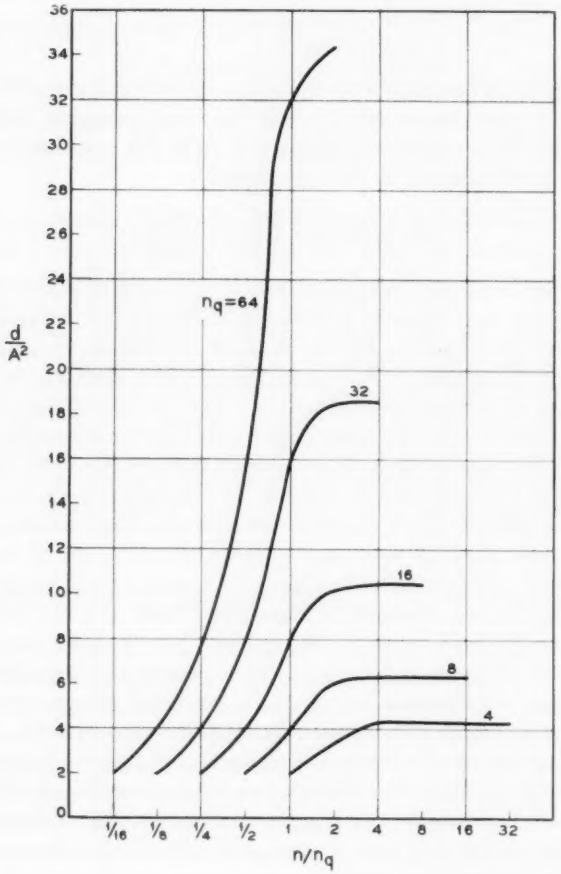


Fig. 3 — Normalized results of a solution of (5) and (6), with d/A^2 as a function of number of samples n/n_q and of stimulus duration in terms of number of independent samples n_q .

longer a factor in detectability since, theoretically at least, any number of derivatives can be measured from as short a sample as desired.

Detectability can again be computed from (5) and (6), where

$$\beta_{rs} = E\langle N^{(r-1)}(0)N^{(s-1)}(0)\rangle$$

is the correlation of the $r - 1$ and $s - 1$ derivatives,

$$N^{(s)}(t) \equiv \frac{d^{s-1}N(t)}{dt^{s-1}}.$$

The correlation coefficient may be written

$$\beta_{rs} = \frac{1}{2\pi} \int_{-\infty}^{+\infty} G(\omega)(-j\omega)^{r-1}(j\omega)^{s-1}\, d\omega, \tag{7}$$

where $G(\omega)$ is the power spectrum of the noise

$$G(\omega) = \int_{-\infty}^{+\infty} E\langle N(t)N(t+\tau)\rangle e^{-j\omega\tau}\, d\tau.$$

If bandlimited noise with a flat spectrum from -1 to $+1$ rad/second and unit rms amplitude is selected, then (7) yields

$$\beta_{rs} = \begin{cases} \left(\dfrac{1}{r+s-1}\right)(-1)^{\frac{1}{2}r+\frac{1}{2}s} & \text{if } r+s \text{ is even} \\ 0 & \text{if } r+s \text{ is odd.} \end{cases}$$

A solution for (5) and (6) with these coefficients can be effected, since the determinants involved are reducible to a form with a solution attributed to Cauchy. The answer can probably be written on a large enough sheet of paper for signals having simple derivatives such as sinusoids, but the result is especially compact for a constant for which

$$S(0) = K, \quad S^{(n)}(0) \equiv \frac{d^{n-1}S(t)}{dt^{n-1}} = 0 \qquad n = 2,3,\cdots.$$

The evaluation, carried out in Appendix A, yields for d

$$d = K^2\left[\frac{(2m)!}{2^{2m-1}m!(m-1)!}\right]^2, \tag{8}$$

where

$$m = \begin{cases} \dfrac{n}{2} & \text{for } n \text{ even} \\ \dfrac{n}{2} + \dfrac{1}{2} & \text{for } n \text{ odd.} \end{cases}$$

The asymptotic behavior of d for large m can be seen by substituting Stirling's approximation

$$a! \approx \sqrt{2\pi}\, \exp\{-a + (\log a)(a + \tfrac{1}{2})\}$$

for the factorials in (8), thus reducing it to

$$d \approx \frac{4K^2}{\pi} e^{1/(6m^2)} m. \tag{9}$$

The approximation is within 2 per cent for $m \geqq 20$.

Equations (8) and (9) exhibit the behavior of a statistic in which d increases linearly with the number of samples, each sample being a derivative. A similar behavior will be shown for rational noise where one term in the detectability depends linearly on the number of derivatives which exist and form part of the statistic. The bandlimited noise differs from the rational noise in that all its derivatives theoretically exist and the detectability can be made, at least theoretically, as good as desired by making m large enough. Obviously, in any practical case, the number of derivatives which can be estimated is limited. In addition, the characterization of the random process as gaussian undoubtedly fails for high enough derivatives.

Equations (8) and (9) are derived only for a signal which is a constant. However, a similar dependence on m would probably occur for sinusoidal signals.

The prominence of derivatives as an effective statistic for both bandlimited and rational noise gives a possible indication why detectability based on equally spaced time samples increases so slowly. These, being uniformly distributed, give poor estimates of derivatives. A more effective distribution might well be n_q independent samples spaced uniformly over the interval and the rest of the samples clustered as closely as possible about two points at each end of the interval. Such arrangement is suggested by statistics for the rational noise case.

III. DETECTION WITH CONTINUOUS SAMPLING

The preceding section discussed the detection of a known signal in bandlimited noise using a finite number of samples of the stimulus as a statistic. In this section we consider the detection of a known signal in gaussian noise using as the statistic a continuous measure of the stimulus over an interval T in length. The noise is now taken to have a rational power spectral density; that is, its power spectrum can be represented at the ratio of two polynomials in ω^2. Such noise can be thought of as resulting from the passage of ideal white gaussian noise through a finite linear lumped-element filter, although it need not actually have been produced in this way. For the purposes of the analysis, it is convenient to think of the situation as shown in Fig. 4. White gaussian noise is passed through a filter whose transfer function is $H(s)$, (Laplace transform of its impulse response) and to this may or may not be added the known signal $S(t)$. T seconds of the combination form the stimulus $Y(t)$. The problem is to decide from an examination of the stimulus whether or not the signal was present.

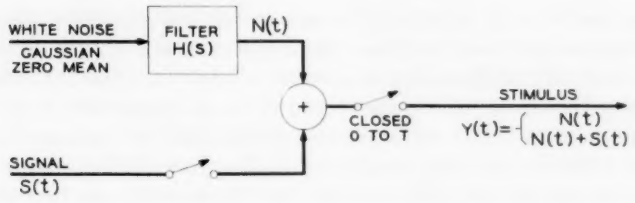


Fig. 4 — Diagram of continuous sampling situation.

The detection scheme in this case is essentially an extension of the finite sampling procedure. One asks for that linear integral operator which will extract from the stimulus a statistic giving the maximum detectability. Thus, the statistic is obtained from

$$\varphi = \int_0^T Y(t)Z(t)\,dt, \tag{10}$$

where $Z(t)$ is that function of time which maximizes the detectability. Because the noise is gaussian of zero mean and the signal (when present) is simply added to the noise, the statistic φ again has a gaussian probability density function whose mean value is zero or not zero according to the absence or presence of the signal and whose variance is the same with or without the signal. Thus it is reasonable to again define the detectability measure d as

$$d = \frac{[E(\varphi_{SN})]^2}{E(\varphi_N{}^2)}. \tag{11}$$

The optimization problem is thus to find $Z(t)$ which maximizes d or, that which is equivalent, to find $Z(t)$ which minimizes $E(\varphi_N{}^2)$ while holding $E(\varphi_{SN})$ constant. This latter form is a straightforward calculus of variation problem and its solution, the details of which are omitted, leads to the following integral equation for $Z(t)$:

$$\int_0^T R(t-u)Z(u)\,du = S(t) \qquad 0 \leqq t \leqq T, \tag{12}$$

where $R(\tau)$ is the autocorrelation function of the noise,

$$R(\tau) = E[N(t)N(t+\tau)].$$

When (12) is satisfied, the detectability can be written

$$d = \int_0^T Z(t)S(t)\,dt. \tag{13}$$

The discussion up to this point has not required that the noise have a rational spectral density. Unfortunately, it does not appear possible to carry (13) any further without actually solving (12) for $Z(t)$, and this has only been done in certain special cases. In particular, if the noise spectral density is the reciprocal of a polynomial, the solution for (12) can be exhibited in some detail; and furthermore if the signal is a sine wave, an exponential, or a constant the detectability can be expressed in a surprisingly simple form.

3.1 *All-Pole Noise*

If the noise has a spectral density $G(\omega)$,

$$G(\omega) = \int_{-\infty}^{+\infty} R(\tau)e^{-j\omega\tau}\, d\tau,$$

which is rational and contains only poles ($2N$ in number), it can be written in the form

$$G(\omega) = \frac{1}{a_0 - a_2\omega^2 + a_4\omega^4 - \cdots \pm a_{2N}\omega^{2N}}. \tag{14}$$

Such a noise could have been produced by passing white noise of unit spectral density through a filter whose transfer function $H(s)$ has N poles,

$$H(s) = \frac{1}{b_0 + b_1 s + b_2 s^2 + \cdots + b_N s^N} = \frac{1}{P(s)}, \tag{15}$$

and the poles can be placed in evidence by writing the denominator polynomial $P(s)$ as

$$P(s) = \sum_{k=0}^{N} b_k s^k = b_N(s - \gamma_1)(s - \gamma_2),\cdots,(s - \gamma_N), \tag{16}$$

where the γ's are (possibly) complex numbers giving the pole locations and each has a negative real part. In terms of $H(s)$, the spectral density can be written

$$G(\omega) = |\, H(j\omega)\,|^2.$$

Thus the noise can be described in a variety of ways—by the constants $a_0, a_2, \cdots, a_{2N}$, or the set $b_0, b_1, \cdots, b_N$, or the pole locations $\gamma_1, \gamma_2, \cdots, \gamma_N$ and one constant b_N, or even the magnitude and phase of the transfer function $H(s)$ for real frequencies. The particular set of parameters to be used will be chosen to simplify the final answer.

One characteristic of N-pole noise is that its first $N-1$ derivatives exist, while the Nth and higher do not. Because of this it is clear that a necessary condition for finite detectability of a signal $S(t)$ is that its first $N-1$ derivatives be continuous in the interval 0 to T. If this condition is not satisfied; that is, if among the $N-1$ derivatives of $S(t)$ a discontinuity occurs, then the detectability is infinite. This is clearly true, because one could simply differentiate the stimulus enough times to produce a step function in the interval and this could always be found by measuring the change in the differentiated stimulus just before and just after the time of the step.

Using this N-pole noise, it is possible to exhibit explicit solutions to (12) and (13). Unfortunately, strictly speaking, (12) does not have a solution unless $S(t)$ and its derivatives up to order $N-1$ satisfy a certain set of boundary conditions (boundaries at 0 and T). If $S(t)$ does not satisfy this set of boundary conditions, and in general for an arbitrary signal it will not, then (12) has a formal solution if $Z(t)$ includes delta functions and their derivatives to order $N-1$ at the end points of the interval (approached from inside the interval). The details of this argument are presented in Appendix B, where it is shown that the solution to (12) is

$$\begin{aligned} Z(t) &= Z_c(t) + \sum_{i=0}^{N-1} [\alpha_i \delta^{(i)}(t) + \beta_i \delta^{(i)}(t-T)], \\ Z_c(t) &= \sum_{k=0}^{N} a_{2k} S^{(2k)}(t), \end{aligned} \tag{17}$$

where the superscript (n) indicates n-fold differentiation with respect to time, and the α's and β's are given by

$$\begin{aligned} \alpha_i &= \sum_{k=i}^{N-1} b_{k+1} U_2^{(k-i)}(0) \\ \beta_i &= \sum_{k=i}^{N-1} (-1)^k b_{k+1} U_1^{(k-i)}(T), \end{aligned} \qquad i = 0,1,2,\cdots,N-1 \tag{18}$$

with

$$U_1(t) = \sum_{k=0}^{N} b_k S^{(k)}(t) \qquad \text{and} \qquad U_2(t) = \sum_{k=0}^{N} (-1)^k b_k S^{(k)}(t).$$

When this $Z(t)$ is substituted in (13), the detectability becomes

$$d = \int_0^T Z_c(t) S(t)\, dt + \sum_{i=0}^{N-1} (-1)^i [\alpha_i S^{(i)}(0) + \beta_i S^{(i)}(T)]. \tag{19}$$

Among the several other ways of writing d, one which is convenient is the following (partly operator notation):

$$d = \int_0^T U_1^2(t)\, dt + \sum_{i=0}^{N-1} S^{(i)}(0)\{[(-1)^i P_i(p)P(-p) + P_i(-p)P(p)]S(t)\}_{t=0}, \tag{20}$$

where

$$P_i(x) = \sum_{k=i}^{N-1} b_{k+1} x^{k-i}$$

and p is the derivative operator d/dt. The derivatives of $S(t)$ and $U(t)$ at 0 and T are to be interpreted as the limit of the value of the derivatives approached from inside the interval.

The form of $Z(t)$ in (17) is quite interesting. The first part contributes a function of time which is similar to the conventional cross-correlation result. One simply multiplies the stimulus by this function and integrates the product. In the second part, the delta functions, when used with (10) to form the statistic, represent evaluating the stimulus and its first $N - 1$ derivatives at the ends of the interval. The derivatives at the ends give information about the stimulus outside the interval. Essentially they allow prediction or estimation of the stimulus outside the interval, and this information is to be added to that from straight cross-correlation. As N becomes larger the noise spectrum drops off faster at high frequencies and more derivatives of the stimulus are used (more derivatives of the noise exist); effectively, the stimulus can be predicted further outside the interval. Usually, this will mean that the signal can be detected better (see examples below).

3.2 *Damped Sinusoidal Signal*

As a particular example, consider the case in which the signal is a damped sine wave of arbitrary phase,

$$S(t) = Ae^{-\alpha t} \sin(\omega t + \Phi) = \bar{A}e^{\lambda t} + \bar{A}^* e^{\lambda^* t}, \tag{21}$$

where

$$\bar{A} = \frac{A}{2j} e^{j\Phi} \qquad \text{and} \qquad \lambda = -\alpha + j\omega.$$

Since the detectability is of primary interest, specific values for the co-

efficients of the delta functions will not be calculated. The details of the calculations are carried out in Appendix C, where it is shown that

$$d = 2\,\mathrm{Re}\left[\bar{A}^2\,\frac{P^2(\lambda)e^{2\lambda T} - P^2(-\lambda)}{2\lambda}\right] + 2\,|\,\bar{A}\,|^2\left[\frac{|\,P(\lambda)\,|^2 e^{(\lambda+\lambda^*)T} - |\,P(-\lambda)\,|^2}{\lambda + \lambda^*}\right]. \tag{22}$$

3.3 *Exponential Signal*

For an exponential signal,

$$S(t) = Ae^{-\alpha t}$$

and the detectability from (22) becomes

$$d = \frac{A^2}{2\alpha}\,[P^2(\alpha) - P^2(-\alpha)\,e^{-2\alpha T}]. \tag{23}$$

With given signal parameters and noise filter, specific values of detectability can be calculated from this expression.

As the number of poles in the noise filter increases, $P^2(-\alpha)/P^2(\alpha) \to 0$, assuming the poles are bounded away from the imaginary axis and that $\alpha > 0$. In this case d becomes

$$d \to A^2P^2(\alpha)/2\alpha.$$

If as the number of poles is increased the DC gain of the filter is kept constant (or allowed to increase), then $P^2(\alpha)$ increases without bound. This can be seen by thinking of $P(\alpha)$ in factored form, which for constant DC gain looks like

$$P(\alpha) = b_0 \prod_{i=1}^{N} \frac{\alpha - \gamma_i}{-\gamma_i},$$

and noting that $|\,(\alpha - \gamma_i)/\gamma_i\,| > 1$. Thus, for fixed signal, more poles mean more detectability. A similar result obtains if $\alpha < 0$.

A noise filter of particular interest is a Butterworth filter, that is, one whose poles are uniformly distributed on a semicircle in the left-half plane. Such a filter gives noise whose spectrum is maximally flat low-pass and approaches ideal bandlimited noise as the number of poles increases. In this case, the approximate behavior of d for large N can

be calculated by taking the poles as smeared out on a semicircle of radius ω_0. Thus,

$$\ln \frac{P^2(\alpha)}{b_0^2} \cong \frac{2N}{\pi} \int_0^{\pi/2} \ln \left[1 + \left(\frac{\alpha}{\omega_0} \right)^2 + 2 \frac{\alpha}{\omega_0} \cos \Phi \right] d\Phi$$

and, therefore,

$$d \cong \frac{A^2}{2\alpha G(0)} B^N, \tag{24}$$

where

$$B = \exp \left\{ \frac{2}{\pi} \int_0^{\pi/2} \ln \left[1 + \left(\frac{\alpha}{\omega_0} \right)^2 + 2 \frac{\alpha}{\omega_0} \cos \Phi \right] d\Phi \right\}.$$

A sketch of B versus α/ω_0 is shown in Fig. 5. Clearly B is greater than one and the detectability grows exponentially for large N.

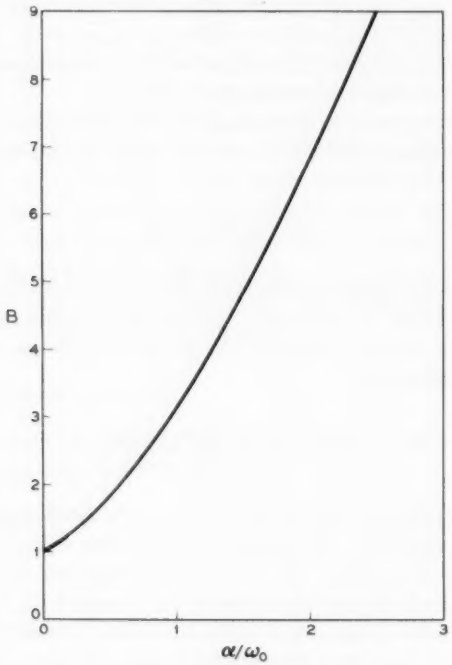


Fig. 5 — B vs. α/ω_0.

3.4 *Sinusoidal Signal*

For an undamped sine wave ($\alpha = 0$), (22) can be put in a more convenient form by using the magnitude and phase of the noise filter transfer function, $H(s)$, which can be written

$$H(j\omega) = \sqrt{G(\omega)}e^{-j\theta(\omega)}.$$

The angle $\theta(\omega)$ then is the phase lag of the noise filter, a function of frequency. In these terms (22) becomes

$$d = \frac{A^2}{2G(\omega)}\left[T + 2\dot{\theta}(\omega) - \frac{\sin(2\omega T + \theta + \Phi) + \sin 2(\theta - \Phi)}{2\omega}\right], \quad (25)$$

where

$$\dot{\theta} = d\theta/d\omega.$$

If $\omega T \gg 1$, that is, if the time is long so that there are many cycles of the sine wave in the interval, then the last term in (25) can be neglected. In conventional circuit analysis, $\dot{\theta}$ is generally considered the time delay of a network; thus, the detectability includes a term proportional to twice the time delay of the noise filter. Roughly, this says that the derivatives at the ends of the interval allow extension of the stimulus a distance equal to the time delay outside each end.

It is clear that the $\dot{\theta}$ term grows without bound as the number of poles bounded away from the imaginary axis is increased. In the particular case of noise with a maximally flat spectrum [Butterworth $H(s)$], this growth can be shown more explicitly. The contribution to $\dot{\theta}$ from a single pair of poles located at $-\omega_0 e^{\pm j\beta}$ is

$$\frac{2}{\omega_0}\frac{(\lambda^2 + 1)\cos\beta}{\lambda^4 + 1 + 2\lambda^2\cos 2\beta}, \qquad \lambda = \frac{\omega}{\omega_0}.$$

To add up the contributions from N poles on a semicircle would lead to a rather complicated expression, but an approximation for large N can be obtained by imagining the poles smeared out on the semicircle, so that the sum can be evaluated as an integral. Then

$$\dot{\theta} \cong \frac{2N}{\pi}\int_0^{\pi/2}\frac{1 + \lambda^2}{\omega_0}\frac{\cos\beta\, d\beta}{1 + \lambda^4 + 2\lambda^2\cos 2\beta} = \frac{2N}{\pi\omega}\ln\left|\frac{\omega + \omega_0}{\omega - \omega_0}\right|. \quad (26)$$

This shows clearly that, for large N, $\dot{\theta}$ increases directly in proportion to N. As a sidelight, the proportionality constant, plotted in Fig. 6, is larger if the signal frequency is near the band edge. The apparent infinity for $\omega = \omega_0$ is a mathematical fiction; it resulted from smearing the poles. For any finite N, $\dot{\theta}$ is finite; thus, the curve in Fig. 6 really should

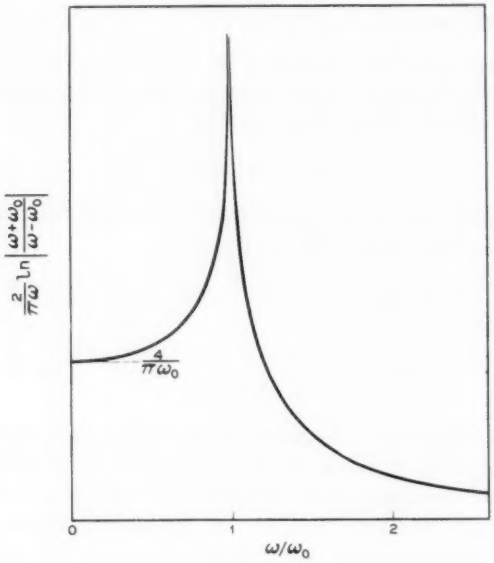


Fig. 6 — Proportionality constant.

be rounded over at the peak. For signal frequencies outside of the noise band, the detectability becomes large simply because the $1/G(\omega)$ term multiplying everything in (25) becomes large. Even straight cross-correlation would give large detectability here.

3.5 *Constant Signal*

For a constant signal, $S(t) = A$, the detectability can be written (see Appendix C)

$$d = \frac{A^2}{G(0)}\left[T - 2\sum_{k=1}^{N}\frac{1}{\gamma_k}\right]. \tag{27}$$

Note that the minus sign does not imply negative detectability; the γ's have negative real parts and so their sum will be negative. Equation (27) shows clearly that the detectability increases as the number of poles bounded away from the imaginary axis is increased.

For N-pole Butterworth noise of bandwidth ω_0, (27) becomes (exactly)

$$d = \frac{A^2}{G(0)}\left[T + \frac{2}{\omega_0 \sin(\pi/2N)}\right],$$

which for large N becomes

$$d \xrightarrow[N\to\infty]{} \frac{A^2}{G(0)}\left[T + \frac{4N}{\pi\omega_0}\right].$$

Here again the detectability grows directly in proportion to N for large N.

IV. CONCLUSIONS

We have presented solutions to some problems involving detection of the presence of known signals in gaussian noise. Thus, we are concerned with what a statistician would term hypothesis testing. Two general classes of detectors are studied, the first a maximum likelihood detector operating on a finite number of samples of the stimulus, the second an optimum integral operator treating the stimulus as a continuous function. However, the new results lie not in the general detection equations, which differ little from ones previously given, but rather in the specific solutions to these equations.

In the finite sampling case, detectability of a sinusoid or constant in bandlimited noise is computed for the cases where the samples are equally spaced time samples spread over a finite duration and where the samples are measurements of successive derivatives at one point in time. As the number of samples increases, detectability increases without bound for both cases. However, for the time samples the rate of increase is very slow for a large number of samples while for derivatives the rate becomes a linear function of the number of samples.

For optimum linear integral detection a general solution is presented for arbitrary signals in noise with a rational all-pole spectrum. The solution in closed form is sufficiently tractable so that the asymptotic behavior of certain simple signals can be evaluated as the number of poles in the noise becomes very large. The solution puts in evidence two different detection mechanisms, one involving integration of the product of the stimulus with a function derived from the signal, the other involving measurement of the derivatives of the stimulus. The first is denoted correlation detection, the second extrapolation detection. Usually, the term arising from correlation detection is the more important. However, if the stimulus is very short or if the noise spectrum has a great number of poles, the extrapolation term may become relatively large. For signals such as a sinusoid it grows without bound as the number of poles increases.

What are the implications of these solutions on previous detection results? Probably they have very little bearing on the perception prob-

lem which engendered the study, since it seems unlikely that animal sense organs embody the mechanisms implied by the solutions or that the characterization of exactly known signals in gaussian noise is appropriate. Both the solutions and the character of the stimuli differ significantly from the Fourier series bandlimited case treated by Peterson, Birdsall and Fox. In particular, the extrapolation detection does not appear in their universe. Also, we feel that the characterization of the noise as described by a correlation function is, to say the least, more suited to the present style of engineering and, to say the most, a much more satisfactory model of most detection situations.

The practical impact, if any, of the detectors developed here would seem to inhere in situations where short pieces of valuable signals must be detected and a great quantity of computing equipment is available. Such might be the case for some space communication problems.

A number of unsolved problems arise directly from the work. For a finite number of time samples of the stimulus, the optimum distribution in time of these samples is unknown. Spectra with zeros as well as poles have not been treated with anything near the elegance of the pure pole situation. Only very specific classes of signals have been studied. It would be of interest to establish which signals give unbounded and which give bounded detectability as the number of poles in the noise increases. Finally, only the case of signals known exactly has been examined. The far more difficult area involving signals with random parameters is almost untouched so far as practical solutions are concerned.

V. ACKNOWLEDGMENT

The authors would like to express their appreciation for the advice and assistance of D. Slepian and J. L. Kelly, Jr., which greatly furthered this work.

APPENDIX A

Detection in Flat Bandlimited Noise by Estimating Derivatives

In the main body of the paper it was shown that, for samples which are derivatives, detectability in terms of d can be determined from (5) and (6)

$$\sum_i \beta_{ij} Z_i = S_j, \qquad j = 1, \cdots, n \tag{5}$$

$$d = \sum_j Z_j S_j . \tag{6}$$

S_j is the $j-1$ derivative of the signal evaluated at $t=0$ and β_{ij}, the correlation coefficient of the noise derivatives, is

$$\beta_{rs} = \begin{cases} \left(\dfrac{1}{r+s-1}\right)(-1)^{\frac{1}{2}r+\frac{1}{2}s} & \text{if } r+s \text{ is even} \\ 0 & \text{if } r+s \text{ is odd} \end{cases}$$

for flat bandlimited noise with unit rms amplitude.

Equation (5) may be written out in matrix form for odd n as

$$\begin{bmatrix} 1 & 0 & -1/3 & 0 & \cdots & 0 & \pm 1/n \\ 0 & 1/3 & 0 & -1/5 & \cdots & \mp 1/n & 0 \\ -1/3 & 0 & 1/5 & 0 & \cdots & & \\ \vdots & & & & & & \\ \pm 1/n & 0 & \cdots & & & & 1/(2n-1) \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \\ \vdots \\ Z_n \end{bmatrix} = \begin{bmatrix} S_1 \\ S_2 \\ S_3 \\ \vdots \\ S_n \end{bmatrix}$$

and a similar form for even n.

This equation may be simplified by separating into two equations and multiplying by minus one in appropriate places to remove minus signs. Two forms occur, one for even n, the other for odd n. For n odd,

$$\begin{bmatrix} 1 & 1/3 & 1/5 & \cdots & 1/n \\ 1/3 & 1/5 & 1/7 & \cdots & \\ 1/5 & 1/7 & 1/9 & \cdots & \\ \vdots & & & & \\ 1/n & \cdots & & & 1/(2n-1) \end{bmatrix} \begin{bmatrix} Z_1 \\ -Z_3 \\ Z_5 \\ \vdots \\ \pm Z_n \end{bmatrix} = \begin{bmatrix} S_1 \\ -S_3 \\ S_5 \\ \vdots \\ \pm S_n \end{bmatrix} \tag{28}$$

and

$$\begin{bmatrix} 1/3 & 1/5 & 1/7 & \cdots & 1/n \\ 1/5 & 1/7 & 1/9 & \cdots & \\ 1/7 & 1/9 & \cdots & & \\ \vdots & & & & \\ 1/n & \cdots & & & 1/(2n-3) \end{bmatrix} \begin{bmatrix} Z_2 \\ -Z_4 \\ Z_6 \\ \vdots \\ \pm Z_{n-1} \end{bmatrix} = \begin{bmatrix} S_2 \\ -S_4 \\ S_6 \\ \vdots \\ \pm S_{n-1} \end{bmatrix}. \tag{29}$$

For n even,

$$\begin{bmatrix} 1 & 1/3 & 1/5 & \cdots & 1/(n-1) \\ 1/3 & 1/5 & 1/7 & \cdots & \\ 1/5 & 1/7 & 1/9 & \cdots & \\ \vdots & & & & \\ 1/(n-1) & & & & (2n-3) \end{bmatrix} \begin{bmatrix} Z_1 \\ -Z_3 \\ Z_5 \\ \vdots \\ \pm Z_{n-1} \end{bmatrix} = \begin{bmatrix} S_1 \\ -S_3 \\ S_5 \\ \vdots \\ \pm S_{n-1} \end{bmatrix} \tag{30}$$

and

$$\begin{bmatrix} 1/3 & 1/5 & 1/7 & \cdots & 1/(n+1) \\ 1/5 & 1/7 & 1/9 & \cdots & \\ 1/7 & 1/9 & \cdots & & \\ \vdots & & & & \\ 1/(n+1) & & & & 1/(2n-1) \end{bmatrix} \begin{bmatrix} Z_2 \\ -Z_4 \\ Z_6 \\ \vdots \\ \pm Z_n \end{bmatrix} = \begin{bmatrix} S_2 \\ -S_4 \\ S_6 \\ \vdots \\ \pm S_n \end{bmatrix}. \tag{31}$$

The determinants of these matrices can be evaluated by applying a rule attributed to Cauchy. In general, the rule says that a determinant whose ijth element is

$$M_{ij} = \frac{1}{a_i + b_j}$$

has the value

$$\left| \frac{1}{a_i + b_j} \right| = \frac{\prod_{j=1}^{n-1} \prod_{i=j+1}^{n} (a_i - a_j)(b_i - b_j)}{\prod_{j=1}^{n} \prod_{i=1}^{n} (a_i + b_j)}.$$

For the particular cases considered here, a_i and b_j have especially simple forms. For example, for (28), $a_i = 2i - 1$ and $b_j = 2j - 2$.

In addition, all cofactors of the matrices are also of Cauchy form. Hence, it is possible to invert the matrices by the method of cofactors and thus solve the equations. Such solutions are quite complex for arbitrary signals. However, an especially simple answer can be obtained for a constant since

$$S_1 = K,$$

$$S_i = 0 \qquad i \neq 1,$$

where K is the signal amplitude. Equation (6) reduces to

$$d = Z_1 S_1 = Z_1 K.$$

Z_1 may be determined by the well-known method for solving equations as the ratio of two determinants,

$$Z_1 = \frac{\begin{vmatrix} K & 1/3 & 1/5 & \cdots & 1/(2m-1) \\ 0 & 1/5 & \cdots & & \\ \vdots & & & & \\ 0 & 1/(2m+1) & & & 1/(4m-3) \end{vmatrix}}{\begin{vmatrix} 1 & 1/3 & 1/5 & \cdots & 1/(2m-1) \\ 1/3 & 1/5 & \cdots & & \\ \vdots & & & & \\ 1/(2m-1) & & & & 1/(4m-3) \end{vmatrix}},$$

where

$$m = \begin{cases} \dfrac{n+1}{2} & \text{for } n \text{ odd} \\ \dfrac{n}{2} & \text{for } n \text{ even.} \end{cases}$$

Application of Cauchy's rule and the solution for d yields

$$d = K^2 \left[\frac{(2m)!}{2^{2m-1} m!\,(m-1)!} \right]^2,$$

which is the result utilized in the main part of the paper.

APPENDIX B

In this appendix we give a general solution to the integral equation

$$\int_0^T R(t-u) Z(u)\, du = S(t) \qquad 0 \leqq t \leqq T, \tag{32}$$

where $R(t)$ is the correlation function of a noise whose spectral density is a rational function of frequency having only poles and $S(t)$ is an arbitrary known signal. The solution of the equation can be expressed in a number of different forms. The particular one developed here has

the great advantage of being an explicit function of $R(t)$ and $S(t)$ rather than involving the solution of a set of linear equations. In addition, it possesses the aesthetically pleasing property of not involving analytic continuation of $S(t)$ outside the interval $0 \leqq t \leqq T$. The noise spectral density can be written

$$G(\omega) = \frac{1}{Q(p)}\bigg|_{p=j\omega}, \qquad Q(p) = \sum_{k=0}^{N} a_{2k}p^{2k}. \tag{33}$$

If we think of $Q(p)$ as an operator with p interpreted as d/dt we see that

$$Q(p)[R(t)] = Q(p)\int_{-\infty}^{+\infty} \frac{1}{Q(j\omega)}\, e^{j\omega t}\, \frac{d\omega}{2\pi} = \delta(t), \tag{34}$$

where $\delta(t)$ is the Dirac delta function. Operating formally on both sides of (32) with $Q(p)$ yields

$$Z_c(t) = Q(p)[S(t)] = \sum_{k=0}^{N} a_{2k}S^{(2k)}(t), \qquad 0 < t < T. \tag{35}$$

The subscript has been added to Z to indicate that this may be only part of the answer and the superscript (n) indicates n-fold differentiation with respect to time. If (32) had a $Z(t)$ solution which was continuous, then (35) would be that solution. But the fact that (35) is continuous (as it would be if $S(t)$ and its derivatives were continuous) does not prove that it is the complete solution. In fact, one can readily verify that (35) is not the complete solution by inserting it back in (32) and seeing if (32) is satisfied. It turns out that (35) is indeed part of the answer, and the remaining part is found by just this process of inserting (35) back in (32) and finding what is missing. If we imagine for the moment that $S(t)$ is extended in some arbitrary way outside the interval (so that it is Fourier transformable and the function and its derivatives go to zero at $\pm\infty$), we can write

$$\int_0^T R(t-u)Z_c(u)\,du = \left[\int_{-\infty}^{+\infty} - \int_{-\infty}^{0} - \int_{T}^{+\infty}\right][R(t-u)Z_c(u)\,du]. \tag{36}$$

The first integral on the right is a normal convolution of Z_c and R, and if Z_c from (35) is substituted in we get back exactly $S(t)$. The second and third integrals are evaluated by repeated partial integration, or,

what is equivalent, by finding an exact differential expression for the integrand. We first note that $Q(p)$ can always be factored,

$$Q(p) = P(p)P(-p), \qquad P(p) = \sum_{k=0}^{N} b_k p^k, \tag{37}$$

where $P(p)$ contains only left-half plane zeros. Now define

$$U_1(t) = P(p)[S(t)] = \sum_{k=0}^{N} b_k S^{(k)}(t)$$

and (38)

$$U_2(t) = P(-p)[S(t)] = \sum_{k=0}^{N} (-1)^k b_k S^{(k)}(t).$$

The exact differential that we need is obtained by clairvoyance. It is

$$\begin{aligned} \frac{d}{du} \sum_{j=1}^{N} \sum_{i=j}^{N} b_i {U_2}^{(i-j)}(u) R^{(j-1)}(t-u) \\ = \sum_{k=0}^{N} b_k [{U_2}^{(k)}(u) R(t-u) - U_2(u) R^{(k)}(t-u)] \qquad (39) \\ = Z_c(u) R(t-u) - U_2(u) \sum_{k=0}^{N} b_k R^{(k)}(t-u). \end{aligned}$$

Now, since $P(p)$ has the left-half plane zeros of $Q(p)$, the Fourier transform of $P(p)[R(t)]$ will have only right-half plane poles and thus

$$P(p)[R(t)] = \sum_{k=0}^{N} b_k R^{(k)}(t) = 0 \qquad \text{for} \quad t > 0.$$

Therefore, when we use (39) in the middle integral on the right of (36), we get

$$\int_{-\infty}^{0} R(t-u) Z_c(u)\, du = \sum_{i=0}^{N-1} \sum_{k=i}^{N-1} b_{k+1} {U_2}^{(k-i)}(0) R^{(i)}(t). \tag{40}$$

The third integral on the right of (36) is evaluated in a similar way, using now (39) with U_2 replaced by U_1 and b_k by $(-1)^k b_k$, and noting that $P(-p)R(t) = 0$ for $t < 0$. In this way we get

$$\int_{T}^{\infty} R(t-u) Z_c(u)\, du = \sum_{i=0}^{N-1} \sum_{k=i}^{N-1} (-1)^k b_{k+1} {U_1}^{(k-i)}(T) R^{(i)}(t-T). \tag{41}$$

It is interesting to note that (40) and (41) depend only on values of $S(t)$ inside the interval $0 \leqq t \leqq T$, so that the way in which $S(t)$ was

extended outside the interval does not matter. To summarize this, we find

$$\int_0^T R(t-u)Z_c(u)\,du = S(t) - \sum_{i=0}^{N-1}\sum_{k=i}^{N-1}[b_{k+1}U_2^{(k-i)}(0)R^{(i)}(t) + (-1)^k b_{k+1}U_1^{(k-i)}(T)R^{(i)}(t-T)]. \tag{42}$$

It is now clear that, for Z_c to be the complete solution to (32), the double sum in (42) must be zero for all t in the interval. This is equivalent to the following boundary conditions on $S(t)$:

$$\begin{aligned} \sum_{k=i}^{N-1} b_{k+1}U_2^{(k-i)}(0) &= 0 \\ \sum_{k=i}^{N-1}(-1)^k b_{k+1}U_1^{(k-i)}(T) &= 0 \end{aligned} \qquad i = 0, 1, \cdots, (N-1). \tag{43}$$

If the signal is such that these conditions are not satisfied, then (32) has a solution only if $Z(t)$ includes delta functions and their derivatives, that is

$$Z(t) = Z_c(t) + \sum_{i=0}^{N-1}[\alpha_i\delta^{(i)}(t) + \beta_i\delta^{(i)}(t-T)]. \tag{44}$$

If this is used in (32), the delta functions bring out R and its derivatives evaluated at t and $t-T$, and the α's and β's can be directly identified as

$$\begin{aligned} \alpha_i &= \sum_{k=i}^{N-1} b_{k+1}U_2^{(k-i)}(0), \\ \beta_i &= \sum_{k=i}^{N-1}(-1)^k b_{k+1}U_1^{(k-i)}(T). \end{aligned} \tag{45}$$

The detectability for a Z which satisfies (32) is

$$d = \int_0^T S(t)Z(t)\,dt;$$

thus, using (44),

$$d = \int_0^T S(t)Z_c(t)\,dt + \sum_{i=0}^{N-1}(-1)^i[\alpha_i S^{(i)}(0) + \beta_i S^{(i)}(T)]. \tag{46}$$

This can be put in a slightly different form which may be more convenient by again partially integrating. Using another exact differential obtained by clairvoyance, which is

$$\frac{d}{dt}\left[\sum_{i=0}^{N-1}\sum_{k=i}^{N-1}(-1)^{k+i}b_{k+1}U_1^{(k-i)}(t)S^{(i)}(t)\right] = -S(t)Z_c(t) + U_1^2(t) \quad (47)$$

and observing that when this is inserted in (46) the terms evaluated at T cancel, we get

$$\begin{aligned} d = &\int_0^T U_1^2(t)\,dt \\ &+ \sum_{i=0}^{N-1}(-1)^i S^{(i)}(0)\sum_{k=i}^{N-1} b_{k+1}[U_2^{(k-i)}(0) + (-1)^k U_1^{(k-i)}(0)] \end{aligned} \quad (48)$$

or, equivalently, in an operator notation,

$$\begin{aligned} d = &\int_0^T U_1^2(t)\,dt \\ &+ \sum_{i=0}^{N-1} S^{(i)}(0)[(-1)^i P_i(p)U_2(t) + P_i(-p)U_1(t)]_{t=0}, \end{aligned} \quad (49)$$

where

$$P_i(x) = \sum_{k=i}^{N-1} b_{k+1}x^{k-i}.$$

In this form the summation only involves derivatives at $t = 0$, which in some cases simplifies the algebra of a solution.

APPENDIX C

As a particular example, we calculate the detectability d for the case in which the signal is an exponentially damped sine wave,

$$S(t) = Ae^{-\alpha t}\sin(\omega t + \Phi) = \bar{A}e^{\lambda t} + \bar{A}^* e^{\lambda^* t}, \quad (50)$$

where

$$\bar{A} = \frac{A}{2j}e^{j\Phi} \qquad \text{and} \qquad \lambda = -\alpha + j\omega$$

and the asterisk denotes complex conjugate. Using this in the expression

for detectability (20) or (49), we find that the second term—call it d_d — becomes

$$\begin{aligned} d_d &= \sum_{i=0}^{N-1} S^{(i)}(0)\{[(-1)^i P_i(p)P(-p) + P_i(-p)P(p)]S(t)\}_{t=0} \\ &= 2\,\mathrm{Re}\sum_{i=0}^{N-1}\{\bar{A}^2[(-\lambda)^i P_i(\lambda)P(-\lambda) + \lambda^i P_i(-\lambda)P(\lambda)] \\ &\quad + |\bar{A}|^2[(-\lambda^*)^i P_i(\lambda)P(-\lambda) + \lambda^{*i} P_i(-\lambda)P(\lambda)]\}. \end{aligned} \tag{51}$$

The notation Re means "real part of." From the definition of $P_i(x)$, one can readily verify that

$$\sum_{i=0}^{N-1} y^i P_i(x) = \frac{P(x) - P(y)}{x - y}, \tag{52}$$

and this allows (51) to be greatly simplified:

$$\begin{aligned} d_d &= 2\,\mathrm{Re}\left[\bar{A}^2\frac{P^2(\lambda) - P^2(-\lambda)}{2\lambda}\right] \\ &\quad + 2|\bar{A}|^2\frac{|P(\lambda)|^2 - |P(-\lambda)|^2}{\lambda + \lambda^*}. \end{aligned} \tag{53}$$

The first term in the detectability, (20) or (49) is simply an integral,

$$\begin{aligned} \int_0^T U_1^2(t)\,dt &= \int_0^T [\bar{A}P(\lambda)e^{\lambda t} + \bar{A}^*P(\lambda^*)e^{\lambda^* t}]^2\,dt \\ &= 2\,\mathrm{Re}\left[\bar{A}^2P^2(\lambda)\frac{e^{2\lambda T} - 1}{2\lambda}\right] + 2|AP(\lambda)|^2\frac{e^{(\lambda+\lambda^*)T} - 1}{\lambda + \lambda^*}. \end{aligned} \tag{54}$$

Combining (53) and (54), we get

$$\begin{aligned} d &= 2\,\mathrm{Re}\left[\bar{A}^2\frac{P^2(\lambda)e^{2\lambda T} - P^2(-\lambda)}{2\lambda}\right] \\ &\quad + 2|\bar{A}|^2\left[\frac{|P(\lambda)|^2e^{(\lambda+\lambda^*)T} - |P(-\lambda)|^2}{\lambda + \lambda^*}\right] \end{aligned} \tag{55}$$

which is the general solution for any damped sinusoid.

Three special cases are now considered, the pure exponential, the pure sine wave, and a constant (DC) signal. For a pure exponential signal, $\lambda \to -\alpha$ where α is real in (55), giving

$$d = \frac{A^2}{2\alpha}[P^2(\alpha) - P^2(-\alpha)e^{-2\alpha T}]. \tag{56}$$

For a pure sine wave signal, $\lambda \to j\omega$. The second term in (55) requires a little special treatment, but it is easily shown that

$$\frac{|P(\lambda)|^2 e^{(\lambda+\lambda^*)T} - |P(-\lambda)|^2}{\lambda + \lambda^*} \xrightarrow[\alpha \to 0]{},$$

$$T|P(j\omega)|^2 + \frac{P(-j\omega)}{j}\frac{dP(j\omega)}{d\omega} - \frac{P(j\omega)}{j}\frac{dP(-j\omega)}{d\omega}.$$

Now, $P(j\omega)$ is simply the reciprocal of the transfer function of the noise filter at the frequency ω; that is,

$$P(j\omega) = 1/H(j\omega) = \frac{e^{j\theta(\omega)}}{\sqrt{G(\omega)}},$$

where $\theta(\omega)$ is the phase lag of the noise filter. Using this expression,

$$d = \frac{A^2}{2G(\omega)}\left[T + 2\dot{\theta}(\omega) - \frac{\sin 2(\omega T + \theta + \Phi) + \sin 2(\theta - \Phi)}{2\omega}\right], \quad (57)$$

where $\dot{\theta} = d\theta/d\omega$.

For a constant signal we can simply take (56) and let $\alpha \to 0$, which gives

$$\begin{aligned} d &= A^2\left[TP^2(0) + \frac{dP^2(\alpha)}{d\alpha}\bigg|_{\alpha=0}\right] \\ &= \frac{A^2}{G(0)}\left[T + 2\frac{b_1}{b_0}\right] = \frac{A^2}{G(0)}\left[T - 2\sum_{k=1}^{N} 1/\gamma_k\right]. \end{aligned} \quad (58)$$

REFERENCES

1. Grenander, U., Stochastic Processes and Statistical Inference, Arkiv för Matematik, **1,** 1950, p. 195.
2. Zadeh, L. A. and Ragazzini, J. R., An Extension of Wiener's Theory of Prediction, J. Appl. Phys., **21,** 1950, p. 645.
3. Reich, E. and Swerling, P., The Detection of a Sine Wave in Gaussian Noise, J. Appl. Phys., **24,** 1953, p. 289.
4. Davenport, W. B., Jr. and Root, W. L., *Introduction to the Theory of Random Signals and Noise*, McGraw-Hill, New York, 1958.
5. Middleton, D., *An Introduction to the Theory of Statistical Communication*, McGraw-Hill, New York, 1960.
6. Peterson, W. W., Birdsall, T. G. and Fox, W. C., The Theory of Signal Detectability, I.R.E. Trans., **PGIT4**, 1954, p. 171.
7. Slepian, D., Some Comments on the Detection of Gaussian Signals in Gaussian Noise, I.R.E. Trans., **IT4**, 1958, p. 65.

G. E. Schindler, Jr., New Editor of B.S.T.J.

G. E. Schindler, Jr., was appointed editor of the Bell System Technical Journal, effective January 1, 1961. Mr. Schindler studied chemical engineering at the Carnegie Institute of Technology, received the bachelor of science degree from the University of Chicago, and received the master of arts degree in English literature and languages from the University of Pittsburgh. After additional graduate work at the University of Chicago, Mr. Schindler joined Bell Telephone Laboratories in 1953. He was editor of the Bell Laboratories Record from 1957 to 1959, and most recently was with the Public Relations department of the A. T. & T. Co.

Resonant Modes in a Maser Interferometer

By A. G. FOX and TINGYE LI

(Manuscript received October 20, 1960)

A theoretical investigation has been undertaken to study diffraction of electromagnetic waves in Fabry-Perot interferometers when they are used as resonators in optical masers. An electronic digital computer was programmed to compute the electromagnetic field across the mirrors of the interferometer where an initially launched wave is reflected back and forth between the mirrors.

It was found that after many reflections a state is reached in which the relative field distribution does not vary from transit to transit and the amplitude of the field decays at an exponential rate. This steady-state field distribution is regarded as a normal mode of the interferometer. Many such normal modes are possible depending upon the initial wave distribution. The lowest-order mode, which has the lowest diffraction loss, has a high intensity at the middle of the mirror and rather low intensities at the edges. Therefore, the diffraction loss is much lower than would be predicted for a uniform plane wave. Curves for field distribution and diffraction loss are given for different mirror geometries and different modes.

Since each mode has a characteristic loss and phase shift per transit, a uniform plane wave which can be resolved into many modes cannot, properly speaking, be resonated in an interferometer. In the usual optical interferometers, the resolution is too poor to resolve the individual mode resonances and the uniform plane wave distribution may be maintained approximately. However, in an oscillating maser, the lowest-order mode should dominate if the mirror spacing is correct for resonance.

A confocal spherical system has also been investigated and the losses are shown to be orders of magnitude less than for plane mirrors.

I. INTRODUCTION

Schawlow and Townes[1] have proposed infrared and optical masers using Fabry-Perot interferometers as resonators. Very recently, Mai-

man[2] and Collins et al.[3] have demonstrated experimentally the feasibility of stimulated optical radiation in ruby. In these experiments two parallel faces of the ruby sample were polished and silvered so as to form an interferometer. The radiation due to stimulated emission resonates in the interferometer and emerges from a partially silvered face as a coherent beam of light.

In a maser using an interferometer for a resonator, a wave leaving one mirror and traveling toward the other will be amplified as it travels through the active medium. At the same time it will lose some power due to scattering by inhomogeneities in the medium. When the wave arrives at the second mirror some power will be lost in reflection due to the finite conductivity of the mirror and some power will be lost by radiation around the edges of the mirror. For oscillation to occur, the total loss in power due to density scattering, diffractive spillover and reflection loss must be less than the power gained by travel through the active medium. Thus diffraction loss is expected to be an important factor, both in determining the start-oscillation condition, and in determining the distribution of energy in the interferometer during oscillation.

While it is common practice to regard a Fabry-Perot interferometer as being simultaneously resonant for uniform plane waves traveling parallel to the axis and at certain discrete angles from the axis, this picture is not adequate for the computation of diffraction loss in a maser. It is true that, when the interferometer is operated as a passive instrument with uniform plane waves continuously supplied from an external source, the internal fields may be essentially those of uniform plane waves. In an oscillating maser where power is supplied only from within the interferometer, the recurring loss of power from the edges of a wave due to diffraction causes a marked departure from uniform amplitude and phase across the mirror.

The purpose of our study is to investigate the effects of diffraction on the electromagnetic field in a Fabry-Perot interferometer in free space. The conclusions can be applied equally well to gaseous or solid state masers provided the interferometer is immersed in the active medium, i.e., there are no side-wall discontinuities.

II. FORMULATION OF THE PROBLEM

2.1 *General Formulation*

Our approach is to consider a propagating wave which is reflected back and forth by two parallel plane mirrors, as shown in Fig. 1(a). [This is

equivalent to the case of a transmission medium comprising a series of collinear identical apertures cut into parallel and equally spaced black (perfectly absorbing) partitions of infinite extent, as in Fig. 1(b).] We assume at first an arbitrary initial field distribution at the first mirror and proceed to compute the field produced at the second mirror as a result of the first transit. The newly calculated field distribution is then used to compute the field produced at the first mirror as a result of the second transit. This computation is repeated over and over again for subsequent successive transits. The questions we have in mind are: (a) whether, after many transits, the relative field distribution approaches a steady state; (b) whether, if a steady-state distribution results, there are any other steady-state solutions; and (c) what the losses associated with these solutions would be. While it is by no means obvious that steady-state solutions (corresponding to normal modes) exist for a system which has no side-wall boundaries, it will be shown that such solutions do indeed exist.*

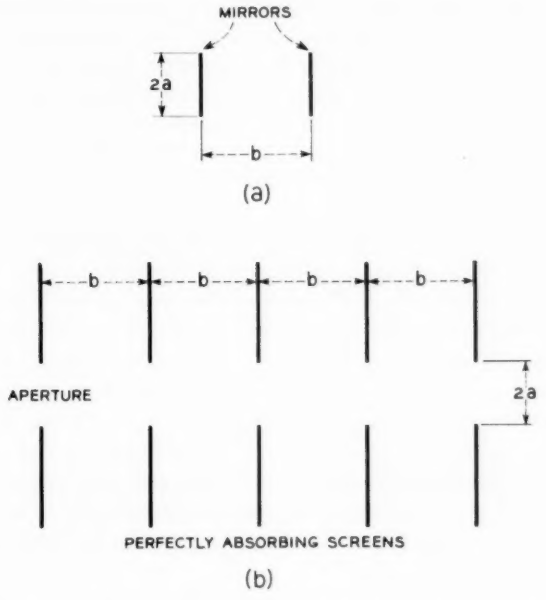


Fig. 1 — The Fabry-Perot interferometer and the transmission medium analog.

* Schawlow and Townes[1] suggested the possibility that resonant modes for a parallel plate interferometer might be similar in form to those for a totally enclosed cavity.

We shall use the scalar formulation of Huygens' principle to compute the electromagnetic field at one of the mirrors in terms of an integral of the field at the other. This is permissible if the dimensions of the mirror are large in terms of wavelength and if the field is very nearly transverse electromagnetic and is uniformly polarized ir one direction. Later, we shall show that these assumptions are consistent with the results of our solutions and therefore are justifiable. We shall also show that other polarization configurations can be constructed from the solutions of the scalar problem by linear superposition.

The Fresnel field u_p due to an illuminated aperture A is given by the surface integral[4]

$$u_p = \frac{jk}{4\pi} \int_A u_a \frac{e^{-jkR}}{R} (1 + \cos \theta)\, dS, \tag{1}$$

where u_a is the aperture field, k is the propagation constant of the medium, R is the distance from a point on the aperture to the point of observation and θ is the angle which R makes with the unit normal to the aperture. We now assume that an initial wave of distribution u_p is launched at one of the mirrors of the interferometer and is allowed to be reflected back and forth in the interferometer. After q transits the field at a mirror due to the reflected field at the other is simply given by (1) with u_p replaced by u_{q+1}, which is the field across the mirror under consideration and u_a by u_q, which is the reflected field across the opposite mirror giving rise to u_{q+1}.

It is conceivable that after many transits the distribution of field at the mirrors will undergo negligible change from reflection to reflection and will eventually settle down to a steady state. At this point the fields across the mirrors become identical except for a complex constant; that is,

$$u_q = \left(\frac{1}{\gamma}\right)^q v\,, \tag{2}$$

where v is a distribution function which does not vary from reflection to reflection and γ is a complex constant independent of position coordinates. Substituting (2) in (1) we have the integral equation

$$v = \gamma \int_A Kv\, dS \tag{3}$$

in which the kernel of the integral equation, K, is equal to $(jk/4\pi R) \cdot (1 + \cos \theta) e^{-jkR}$. The distribution function v, which satisfies (3), can

be regarded as a normal mode of the interferometer defined at the mirror surface, and the logarithm of γ, which specifies the attenuation and the phase shift the wave suffers during each transit, can be regarded as the propagation constant associated with the normal mode.

The integral equation (3) can be solved numerically by the method of successive approximations (Ref. 5, p. 421). It is interesting to note that this iterative method of solution is analogous to the physical process of launching an initial distribution of wavefront in the interferometer and letting it bounce back and forth between the mirrors as described in the foregoing paragraphs.

We have studied and obtained numerical solutions for several geometric configurations of the interferometer. These are (a) rectangular plane mirrors, (b) circular plane mirrors and (c) confocal spherical or paraboloidal mirrors.

2.2 *Rectangular Plane Mirrors*

When the mirror separation is very much larger than the mirror dimensions the problem of the rectangular mirrors reduces to a two-dimensional problem of infinite strip mirrors. This is shown in Appendix A. The integral equation for the problem of infinite strip mirrors, when $a^2/b\lambda$ is much less than $(b/a)^2$, is

$$v(x_2) = \gamma \int_{-a}^{a} K(x_2, x_1) v(x_1)\, dx_1 \tag{4}$$

with

$$K(x_2, x_1) = \frac{e^{j(\pi/4)}}{\sqrt{\lambda b}}\, e^{-jk(x_1 - x_2)^2/2b} \tag{4a}$$

The various symbols are defined in Fig. 2 and Appendix A.

Equation (4) is a homogeneous linear integral equation of the second kind. Since the kernel is continuous and symmetric [$K(x_2, x_1) = K(x_1, x_2)$], its eigenfunctions v_n corresponding to distinct eigenvalues γ_n are orthogonal in the interval $(-a,a)$; that is (Ref. 5, p. 413),

$$\int_{-a}^{a} v_m(x) v_n(x)\, dx = 0, \qquad m \neq n. \tag{5}$$

It should be noted that the eigenfunctions are in general complex and are defined over the surface of the mirrors only. They are not orthogonal in the power (Hermitian) sense as commonly encountered in lossless systems. Here, the system is basically a lossy one and the orthogonality relation is

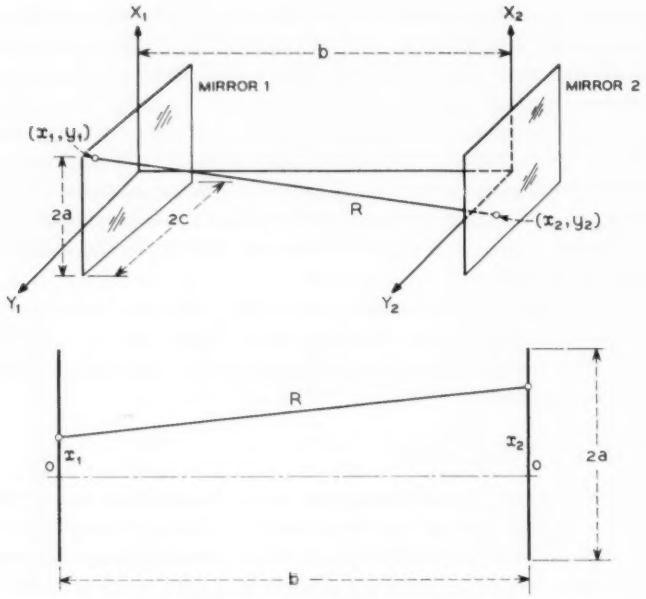


Fig. 2 — Geometry of rectangular plane mirrors.

one which is generally applicable to lossy systems, such as lossy-wall waveguides.

The eigenfunctions are distribution functions of the field over mirror surfaces and represent the various normal modes of the system. The normal modes for rectangular plane mirrors are obtained by taking the products of the normal modes for infinite strip mirrors in x and y directions; that is,

$$v_{mn}(x,y) = v_{x,m}(x)v_{y,n}(y). \tag{6}$$

We designate this as the TEM_{mn} mode for the rectangular plane-mirror interferometer. In view of (5) we see that the normal mode distribution functions v_{mn} are orthogonal over the surface of the rectangular mirror.

The logarithms of the eigenvalues represent propagation constants associated with the normal modes. The propagation constant for the TEM_{mn} mode of rectangular plane mirrors is given by

$$\log \gamma_{mn} = \log \gamma_{x,m} + \log \gamma_{y,n}\,. \tag{7}$$

The real part of the propagation constant specifies the loss per transit

and the imaginary part the phase shift per transit, in addition to the geometrical phase shift, for the normal modes.

2.3 *Circular Plane Mirrors*

It is shown in Appendix B that the solutions to the integral equation for circular plane mirrors (Fig. 3) when $a^2/b\lambda$ is much less than $(b/a)^2$, are given by

$$v(r,\varphi) = R_n(r)e^{-jn\varphi} \qquad (n = \text{integer}), \tag{8}$$

where $R_n(r)$ satisfies the reduced integral equation

$$R_n(r_2)\sqrt{r_2} = \gamma_n \int_0^a K_n(r_2, r_1)R_n(r_1)\sqrt{r_1}\, dr_1, \tag{9}$$

with

$$K_n(r_2, r_1) = j^{n+1}\frac{k}{b} J_n\left(k\frac{r_1 r_2}{b}\right)\sqrt{r_1 r_2}\, e^{-jk(r_1^2+r_2^2)/2b}, \tag{9a}$$

where J_n is a Bessel function of the first kind and nth order. As in the

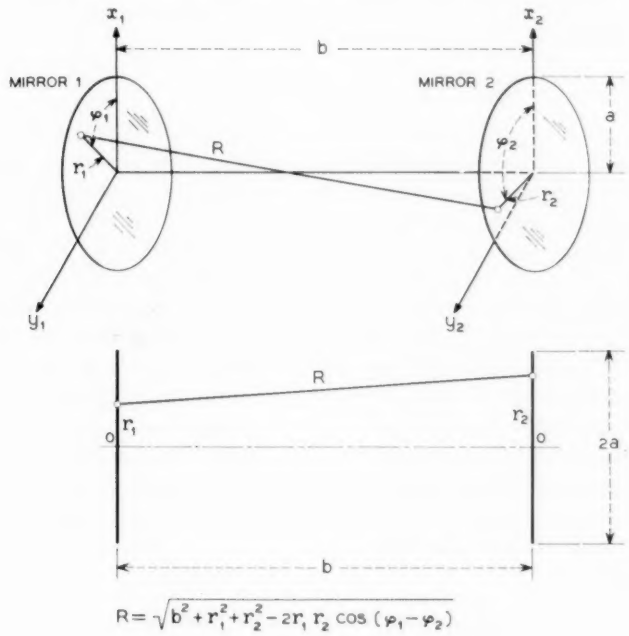


Fig. 3 — Geometry of circular plane mirrors.

problem of infinite strip mirrors, (9) is a homogeneous linear integral equation of the second kind with a continuous and symmetric kernel. Its eigenfunctions corresponding to distinct eigenvalues are orthogonal in the interval $(0,a)$; that is,

$$\int_0^a R_{nl}(r)R_{nm}(r)r\,dr = 0, \qquad (l \neq m). \tag{10}$$

Therefore, we see that the distribution functions $v_{nm}(r,\varphi) = R_{nm}(r)e^{-jn\varphi}$ corresponding to distinct eigenvalues γ_{nm} are orthogonal over the surface of the mirror; that is,

$$\int_0^{2\pi}\int_0^a v_{nm}(r,\varphi)v_{kl}(r,\varphi)r\,dr\,d\varphi = 0 \qquad (\text{either } n \neq k \text{ or } m \neq l). \tag{11}$$

The set of eigenfunctions R_{nm} describes the radial variations of field intensity on the circular mirrors, and the angular variations are sinusoidal in form. We designate a normal mode of the circular plane mirrors as the TEM_{nm} mode, with n denoting the order of angular variation and m denoting the order of radial variation. The propagation constant associated with the TEM_{nm} mode is simply $\log \gamma_{nm}$, which must be obtained from the solution of (9).

2.4 *Confocal Spherical or Paraboloidal Mirrors*

A number of geometries other than plane parallel mirrors have been suggested, and it is believed that most of these can be studied using the same iterative technique. One of the geometries we investigated is that of a confocal spherical system.[6] In this geometry the spherical mirrors have identical curvatures and their foci are coincident, as shown in Fig. 4. One of the possible advantages of such a system is the relative ease of adjustment, since the mirrors are no longer required to be parallel as in the case of the parallel plane system. Another is that the focusing action of the mirrors might give rise to lower diffraction losses.

A spherical mirror with a small curvature approximates closely a paraboloidal mirror. In the case of confocal spherical mirrors, the conditions that its curvature be small is equivalent to saying that the separation between mirrors is large compared to the dimensions of the mirrors. It is shown in Appendix C that the solutions to the integral equation for confocal paraboloidal mirrors, when $a^2/b\lambda$ is much less than $(b/a)^2$, are given by

$$v(r,\varphi) = S_n(r)e^{-jn\varphi} \qquad (n = \text{integer}), \tag{12}$$

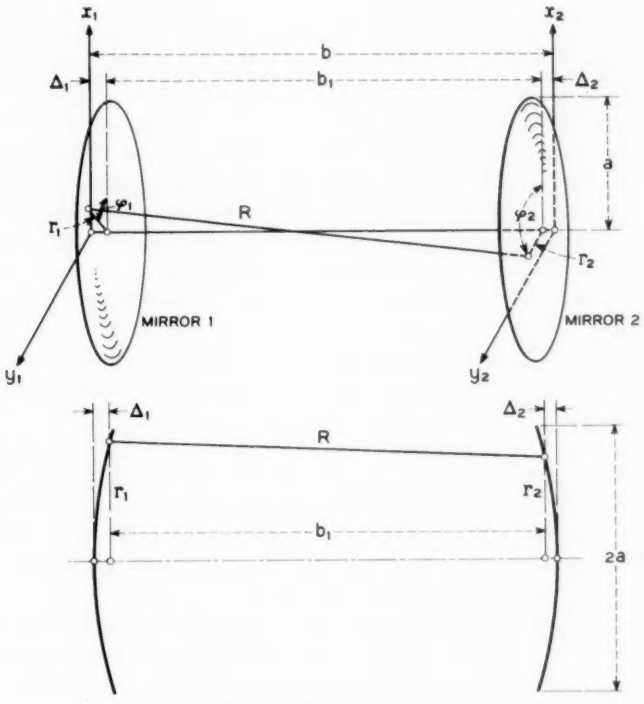
where $S_n(r)$ satisfies the reduced integral equation

$$S_n(r_2)\sqrt{r_2} = \gamma_n \int_0^a K_n(r_2, r_1) S_n(r_1)\sqrt{r_1}\, dr_1, \tag{13}$$

with

$$K_n(r_2, r_1) = j^{n+1} \frac{k}{b} J_n\left(k \frac{r_1 r_2}{b}\right) \sqrt{r_1 r_2}. \tag{13a}$$

Again, we see that (13) is a homogeneous linear integral equation of the second kind with a continuous and symmetric kernel. Therefore, general remarks concerning the normal modes of circular plane mirrors given in the foregoing section are also applicable to confocal spherical or paraboloidal mirrors.



$$R = \sqrt{b_1^2 + r_1^2 + r_2^2 - 2r_1 r_2 \cos(\varphi_1 - \varphi_2)}$$

$$\Delta_1 = b - \sqrt{b^2 - r_1^2}$$

$$\Delta_2 = b - \sqrt{b^2 - r_2^2}$$

Fig. 4 — Geometry of confocal spherical mirrors.

III. COMPUTER SOLUTIONS

3.1 *General*

An IBM 704 computer was programmed to solve the integral equations for the various geometries of the interferometer by the method of successive approximations. As mentioned previously, this is analogous to the physical process of launching an initial distribution of wavefront in the interferometer and letting it bounce to and fro between the mirrors.

3.2 *Infinite Strip Mirrors*

The first problem put on the computer was that of a pair of infinite strip mirrors, having the dimensions $2a = 50\lambda$, $b = 100\lambda$. Equation (26) was employed for the computation, using an initial excitation of a uniform plane wave at the first mirror. A total of one hundred increments was used for the numerical integration. After the first transit the field intensity (electric or magnetic) had the amplitude and phase shown in Fig. 5. In these and subsequent amplitude and phase distributions the curves are normalized so that the maximum amplitude is unity, and the phase at that point is zero. The large ripples are due to the fact that the initial wave front contains 6.25 Fresnel zones as seen from the center of the second mirror. Therefore, in passing from the center to the edge of the second mirror there is a change of 3×6.25 Fresnel zones, and this agrees with the number of reversals in curvature seen in the amplitude distribution.

With subsequent transits, these ripples grow smaller, the amplitude at the edge of the mirror decreases, and the relative field distributions approach a steady state. By the time the wave had made three hundred bounces, the fluctuations occurring from bounce to bounce were less than 0.03 per cent of the final average value. The amplitude and phase for the 300th bounce are also shown in Fig. 5.

We regard this field distribution as an iterative normal mode of the interferometer. In other words, if this distribution is introduced as an initial wave at one mirror it will reproduce the same distribution at the other mirror. Indeed, this is what the computer is verifying when we compute the 301st bounce.

Once the solutions have reached a steady state, we can pick any point on the wavefront, say the center of the mirror, and examine how the absolute phase and amplitude change from bounce to bounce. In this way we determined that the power loss of this mode is 0.688 per cent per transit and the phase shift per transit has a lead of 1.59 degrees.

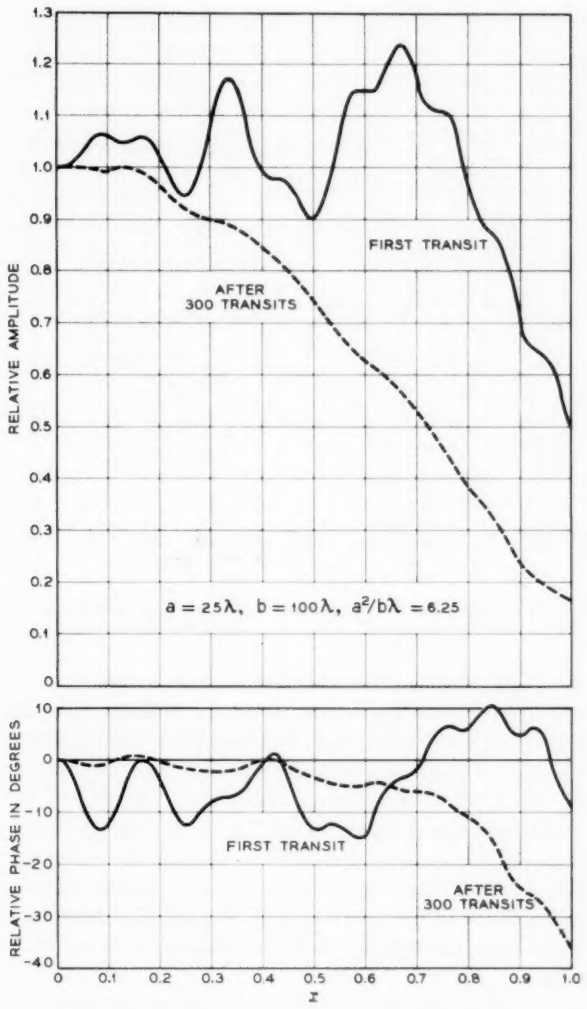


Fig. 5 — Relative amplitude and phase distributions of field intensity for infinite strip mirrors. (The initially launched wave has a uniform distribution.)

Since phase shift is measured relative to the free-space electrical length between the mirrors ($360\ b/\lambda$ degrees), this means that the mode has an effective phase velocity which is slightly greater than the speed of light, just as for a metal tube waveguide.

In Fig. 6 is shown how the field intensity at an arbitrary off-center

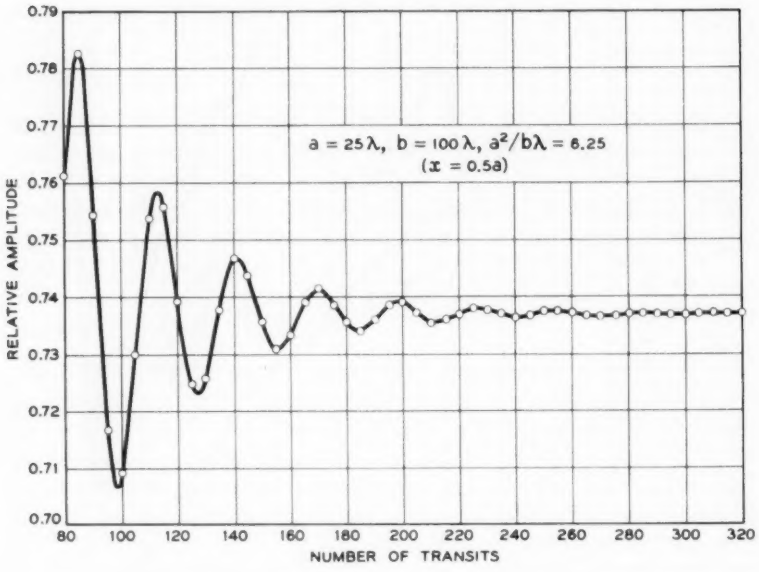


Fig. 6 — Fluctuation of field amplitude at $x = 0.5$ as a function of number of transits. (The initially launched wave has a uniform distribution.)

point ($x = 0.5a$) approaches its steady-state normalized value after a start from a uniform plane wave. After the 100th transit the plot appears to be a damped sine wave. We interpret this damped oscillation as the beating between two normal modes having different phase velocities. The mode with the lower attenuation, of course, survives the longest, and this is the one shown in Fig. 5. We regard this as the dominant mode of the interferometer. We believe the other mode which beats with the dominant mode to be the next-higher order, even-symmetric mode. Prior to the 100th transit, the curve is irregular, indicating that a number of still higher order modes are present which are damped out rapidly.

The next step in the infinite strip problem was to repeat solutions of the above type for other sets of dimensions. However, if $a^2/b\lambda$ is very small compared to $(b/a)^2$, the actual dimensions of the mirrors and their spacing are no longer important, the only parameter of importance being the Fresnel number $N = a^2/b\lambda$. This is approximately equal to the number of Fresnel zones seen in one mirror from the center of the other mirror, and as pointed out earlier, it determines the number of ripples in the field distributions. Amplitude and phase distributions for the

dominant mode obtained by solving (27) are shown in Fig. 7 for different values of N. The larger the N, the weaker is the field intensity at the edge of the mirror, and the smaller is the power loss due to spill-over. The plot of power loss per transit as a function of N is approximately a straight line on log-log paper and is shown as the lowest line in Fig. 8. The phase shift per transit as a function of N is given by the lowest line in Fig. 9.

A uniform plane wave excitation can never give rise to a mode with odd symmetry. In order to investigate the possibility of modes of this

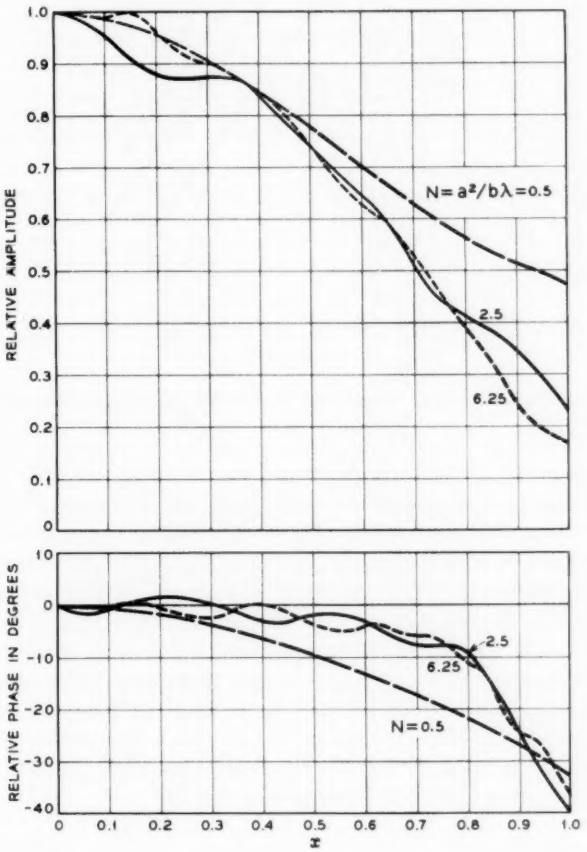


Fig. 7 — Relative amplitude and phase distributions of field intensity of the lowest order even-symmetric mode for infinite strip mirrors.

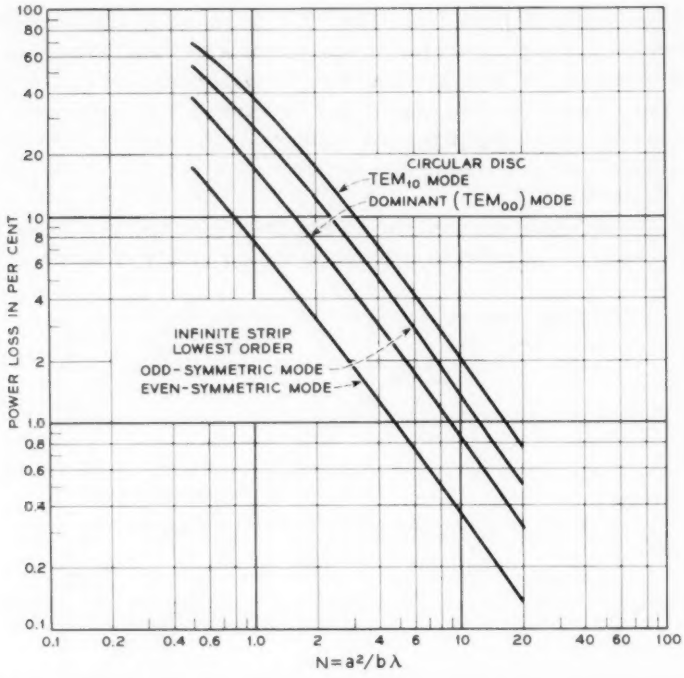


Fig. 8 — Power loss per transit vs. $N = a^2/b\lambda$ for infinite strip and circular plane mirrors.

type, the problem was re-programmed for an initial wave for which the field intensity over one-half the strip (0 to $+a$) was equal but opposite in sign to the field intensity over the other half of the strip (0 to $-a$). Steady-state solutions did indeed result, and odd-symmetric normal modes therefore exist. The amplitude and phase distributions are shown in Fig. 10 for several values of N. The amplitude is zero at the center, as expected. While shown for only one half of the strip, it is the same in the other half, but with a reversal in sign. Note that for the same values of N, the amplitude at the edge is higher than for the dominant mode. The spill-over loss should be higher and this is confirmed by the loss curve in Fig. 8 labeled "infinite strip odd-symmetric mode." The corresponding phase shift curve is shown in Fig. 9.

3.3 *Circular Plane Mirrors*

The feasibility of obtaining the normal mode solutions for the infinite strip mirrors having been established, programs were next set up to

investigate the modes for plane circular mirrors. The first case considered was that for uniform plane wave excitation of the system. Once again, the polarization was assumed to be everywhere parallel to the same axis, and this results in a scalar wave solution having circular symmetry [(9) with $n = 0$]. That is, the amplitude and phase of the field intensity is the same for all points at the same radius from the center. The transverse field distributions for the lowest order mode of this type are shown in Fig. 11 for various values of N. The loss and phase shift are shown in Figs. 8 and 9 under the title "circular disc (dominant mode, TEM_{00})." One hundred increments along the radius were used for the numerical integrations involved.

Next we examined modes of the odd-symmetric type for circular plane mirrors. The equation we used was (9) with $n = 1$. Fig. 12 shows amplitude and phase distributions for the lowest order mode of the odd-symmetric type for circular plane mirrors. Again the loss and phase

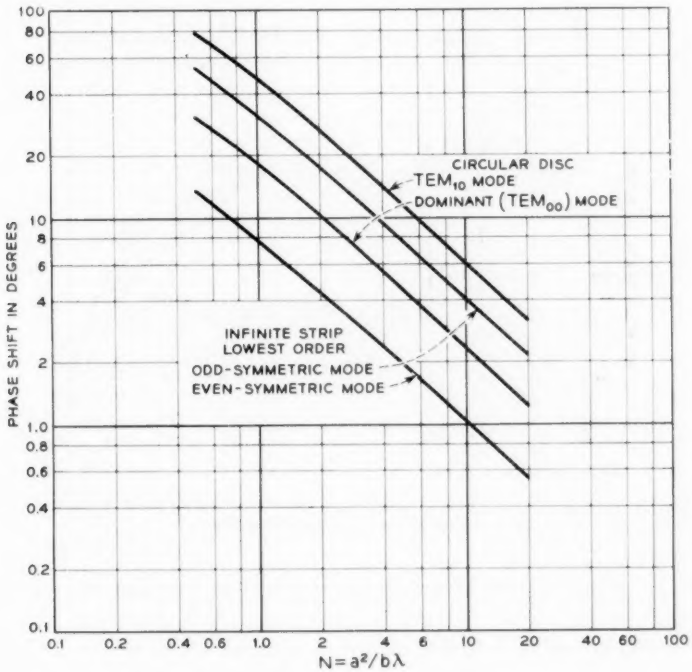


Fig. 9 — Phase shift per transit (leading relative to geometrical phase shift) vs. $N = a^2/b\lambda$ for infinite strip and circular plane mirrors.

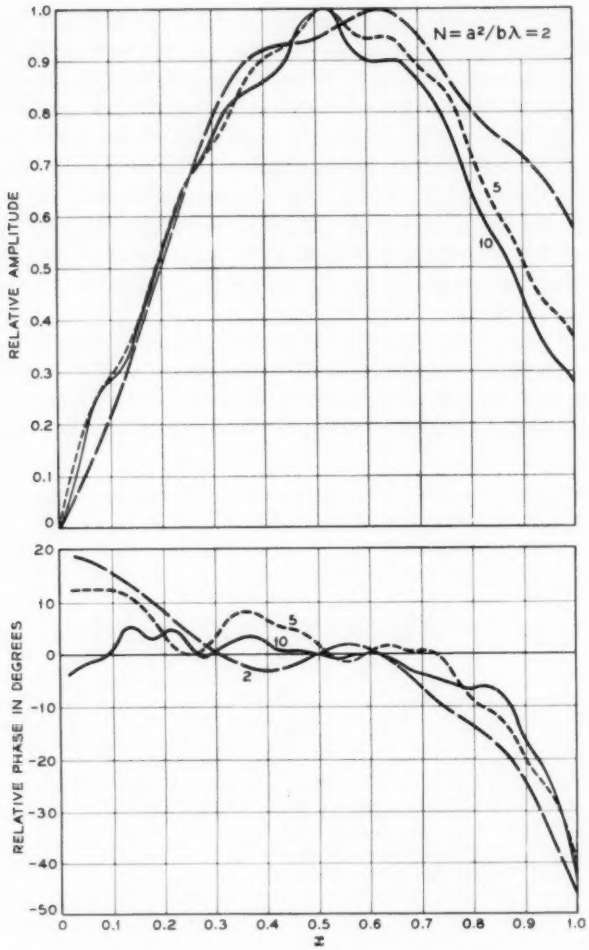


Fig. 10 — Relative amplitude and phase distributions of field intensity of the lowest-order odd-symmetric mode for infinite strip mirrors.

shift are given in Figs. 8 and 9 under the title "circular disc, TEM_{10} mode."

Normal modes with higher orders of angular variation ($n \geqq 2$) and radial variation ($m \geqq 1$) have greater losses and phase shifts than those of TEM_{00} and TEM_{10} modes. The mode with the least attenuation is

therefore the lowest order, of TEM_{00} mode, which we designate as the dominant mode for circular plane mirrors.

3.4 *Confocal Spherical Mirrors*

Before (13) was programmed for solutions on the computer a more general method for solving the problem of the confocal spherical mirrors

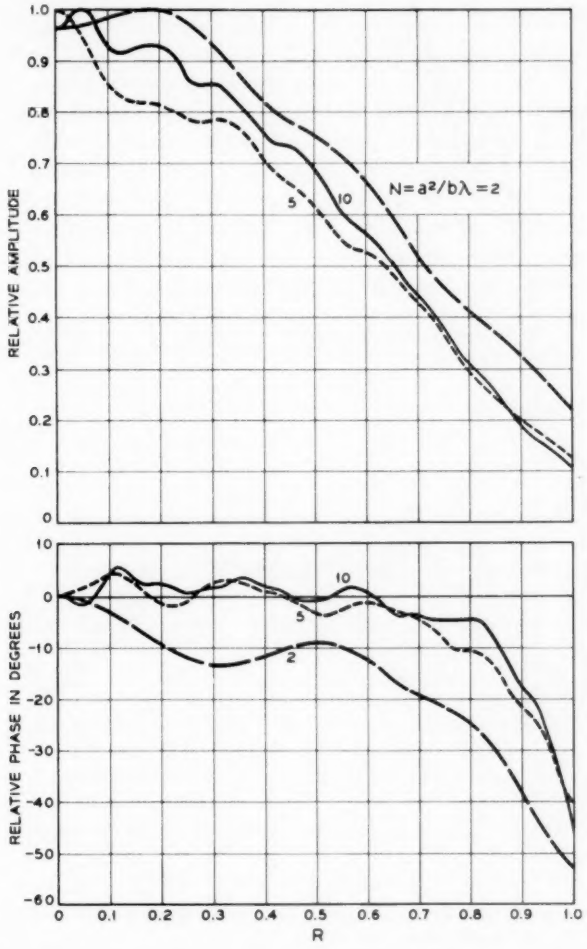


Fig. 11 — Relative amplitude and phase distributions of field intensity of the dominant (TEM_{00}) mode for circular plane mirrors.

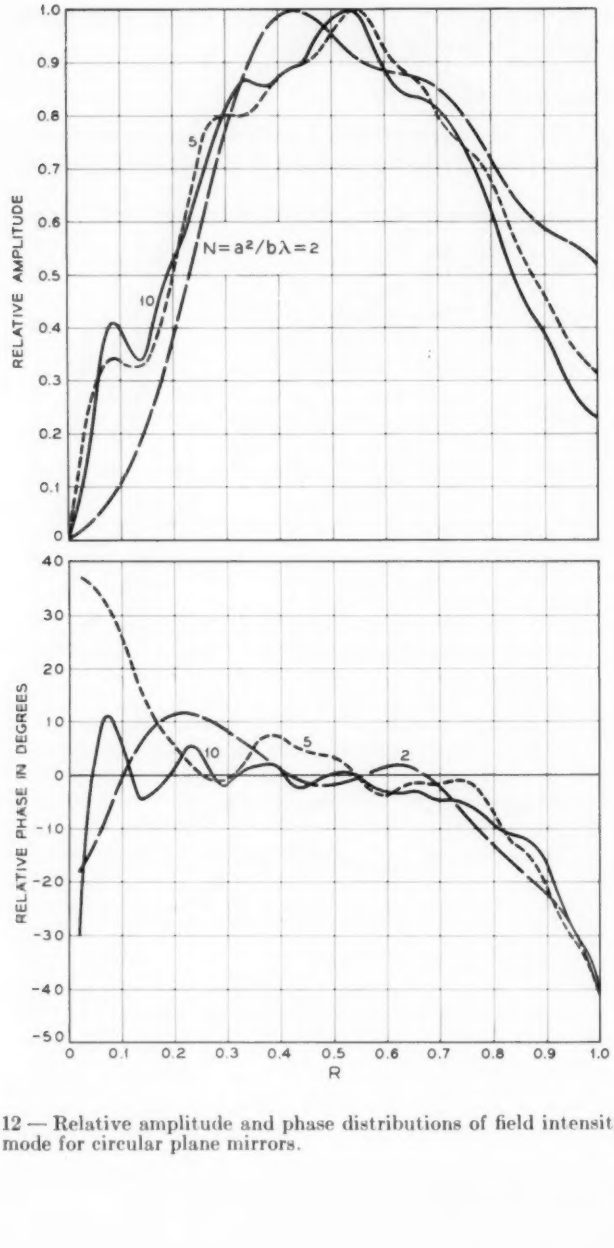


Fig. 12 — Relative amplitude and phase distributions of field intensity of the TEM_{10} mode for circular plane mirrors.

was tried — a procedure that can be used to solve problems involving mirrors with rather arbitrary but small curvatures. In this method the field at each mirror is calculated using the equation for circular plane mirrors and then a phase distribution corresponding to the curvature of the mirror is added to this field before it is used in the next iterative computation. The results from this general method of solution and from solving (13) are in perfect agreement.

The problem of confocal spherical mirrors has also been solved by Goubau[8] and Boyd and Gordon.[9] The results of their analyses are in good agreement with our computed results.

Amplitude distributions of the field intensity for TEM_{00} and TEM_{10} modes are shown in Figs. 13 and 14. The phase distributions are all uniform over the surface of the mirrors and therefore are not plotted. The loss and phase shift per transit are given in Figs. 15 and 16. We note some rather remarkable differences between these solutions and those obtained for circular plane mirrors. First, the field is much more tightly concentrated near the axis of the reflector and falls to a much lower value at the edge than is true for plane mirrors; also the amplitude distribution does not have ripples in it, but is smooth. Second,

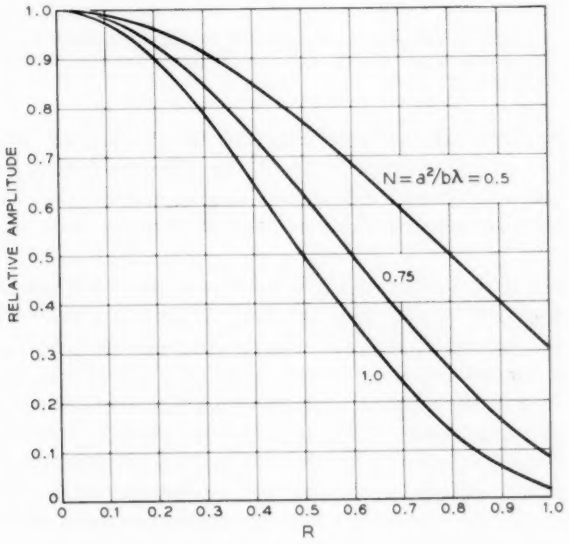


Fig. 13 — Relative amplitude distribution of field intensity of the dominant (TEM_{00}) mode for confocal spherical mirrors. The relative phase distribution on the surface of the mirror is uniform.

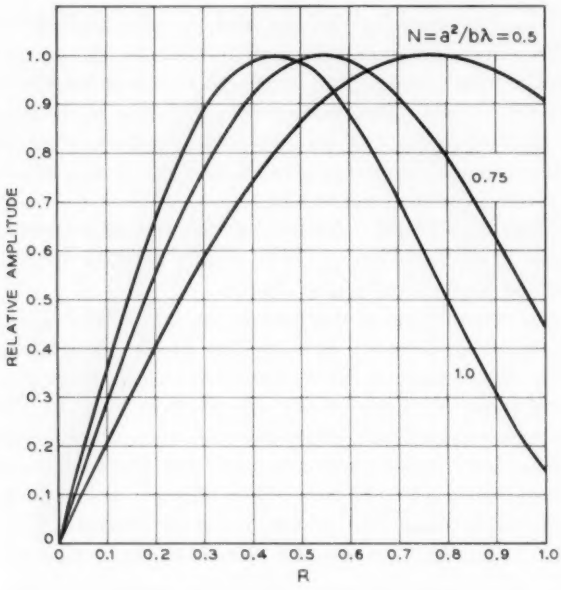


Fig. 14 — Relative amplitude distribution of field intensity of the TEM_{10} mode for confocal spherical mirrors. The relative phase distribution on the surface of the mirror is uniform.

the surface of the reflector coincides with the phase front of the wave, making it an equiphase surface. Third, the difference between the phase shifts for all the normal modes are integral multiples of 90 degrees. Fourth, the losses may be orders of magnitude less than those for plane mirrors.

The result that the mirror surface is an equiphase surface should not be surprising, but can be deduced from integral equation (13). If we associate the factor j^{n+1} with γ_n the kernel becomes real. Since the eigenvalues and eigenfunctions of a real symmetric kernel are all real,[7] we see that the field distribution is of uniform phase over the surface of the mirror. Furthermore, since $(j^{n+1}\gamma_n)$ is real, the phase shift for the normal modes belonging to a set of modes with a given angular variation must be an integral multiple of 180 degrees and the difference between the phase shifts for the normal modes with different angular variations but the same radial variation is an integral multiple of 90 degrees; that is, the phase shift is equal to $[180m + 90(n + 1)]$ degrees. Therefore, if the mirrors are adjusted for the resonance of a particular normal mode,

half of the totality of all the modes are also resonant. However, the resonant mode with the lowest loss would persist longest in the resonator. Just as in the case of plane parallel mirrors, the mode with the lowest loss is the TEM_{00} mode.

IV. DISCUSSION OF RESULTS

The results of machine computation have shown that a two-mirror interferometer, whether of the plane or concave mirror type, can have

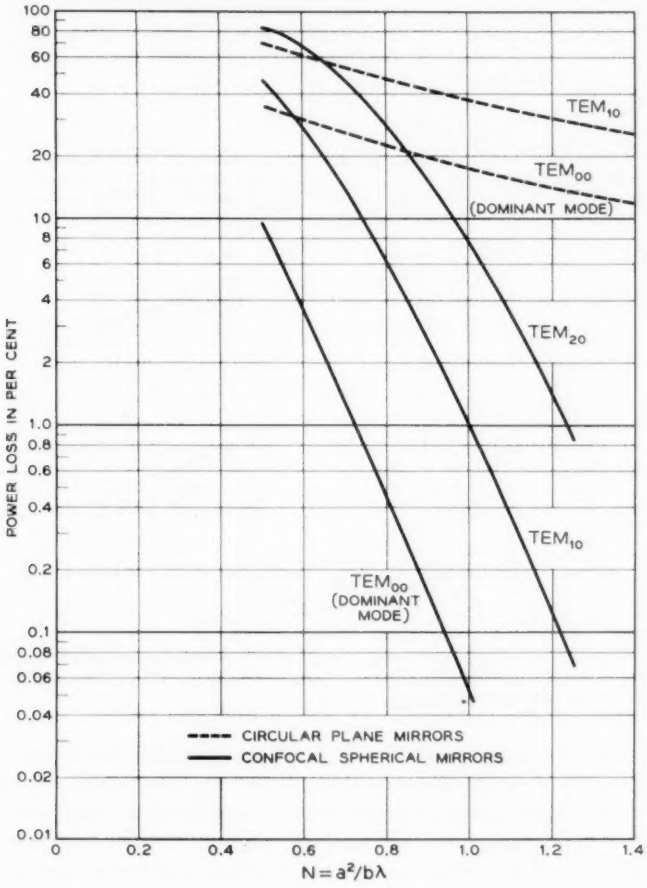


Fig. 15 — Power loss per transit vs. $N = a^2/b\lambda$ for confocal spherical mirrors. (Dashed curves for circular plane mirrors are shown for comparison).

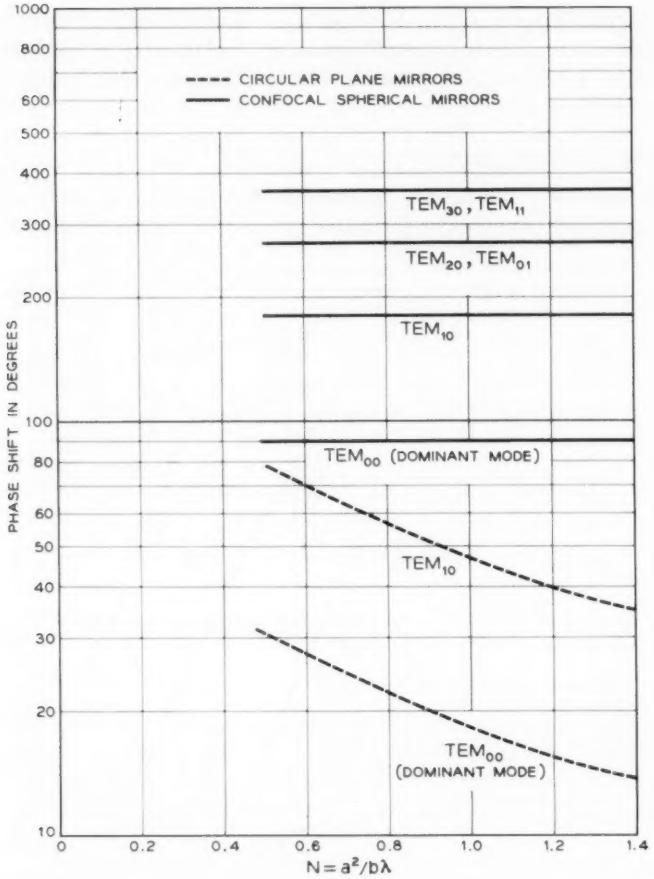


Fig. 16 — Phase shift per transit (leading relative to geometrical phase shift) vs. $N = a^2/b\lambda$ for confocal spherical mirrors. (Dashed curves for circular plane mirrors are shown for comparison.)

normal modes of propagation which are self-perpetuating or self-reproducing in the distance of one transit. We use the term mode of propagation rather than mode of resonance to emphasize the fact that these steady-state solutions are the result of multiple transits whether or not the plate separation happens to be adjusted for resonance. An analog of the plane mirror interferometer is a transmission medium consisting of a series of periodic collinear apertures, as was shown in Fig. 1. The same

solutions apply, and here it is clear that the reproduction of a normal mode field at successive apertures does not depend on any critical relation between b and λ.

In Fig. 17 is shown the way in which a number of square-plate modes can be synthesized from the infinite-strip modes. Diagram A shows schematically the field distribution for the dominant square-plate mode obtained as the product of the field distributions of two even-symmetric strip modes crossed at right angles and with polarization as shown. Since the eigenvalue for the square plate is the product of the eigenvalues for the two strips, the phase shift per transit is the sum of the

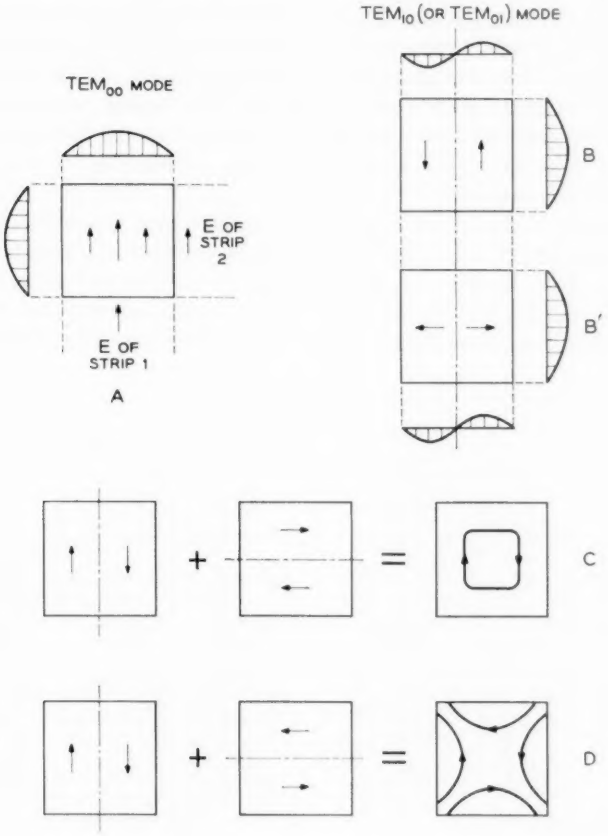


Fig. 17 — Synthesis of normal modes for square mirrors.

phase shifts for the two strips and, if the loss is small, the loss per transit is essentially the sum of the losses for the two strips. Diagram B represents an odd-symmetric square-plate mode formed by taking the product of an even- and an odd-symmetric strip mode; B′ is the same mode but with the polarization rotated 90°; C is a circular electric type of mode formed by *adding* two modes of the type B. This addition is permitted because the two components are degenerate. It follows that the circular electric mode C is degenerate with B and has the same loss and phase shift per transit. By taking the difference between the same two B modes as shown, the mode D is obtained, resembling the TE_{21} mode in circular waveguide. We give all the patterns B, B′, C and D the same designation, TEM_{10} (or TEM_{01}), since they are composites of the one basic mode type. Similar syntheses can be performed for circular mirrors, either plane or concave. It is interesting that degeneracies of this type are common for the interferometer because the electric vector E is at liberty to be parallel or perpendicular to the mirror edges. In a metal waveguide they are uncommon because the polarization of E at the boundaries is restricted.

The dominant mode and a number of higher-order modes for square and circular mirrors are depicted in Fig. 18, in which electric field vectors are shown. This classification of modes applies to plane as well as confocal spherical mirrors. In the case of rectangular mirrors, the x axis may be taken along the longer dimension, in which case the first sub-

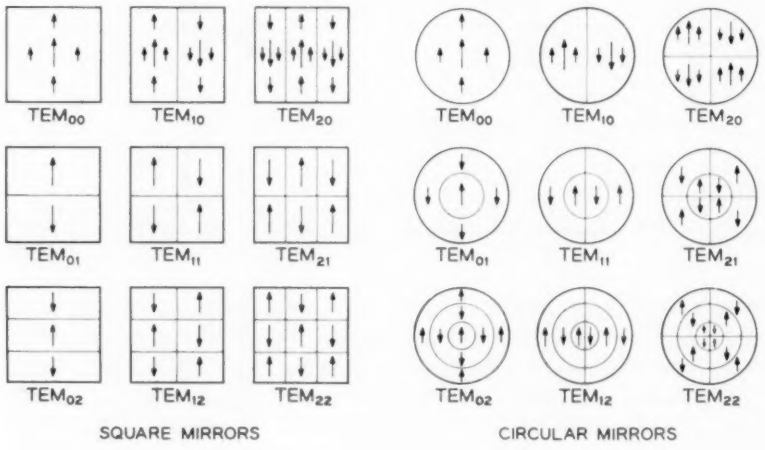


Fig. 18 — Field configuration of normal modes for square and circular mirrors.

script always denotes the number of field reversals along the longer dimension.

In formulating the problem we have assumed that the waves were almost transverse electromagnetic. The solutions for the flat mirror are consistent with this assumption. At the edges of the mirror there is a phase lag of approximately 45 degrees relative to the center, but this is only one-eighth of a wavelength out of many wavelengths for the mirror diameter. Thus the curvature of the wavefront away from the transverse plane is exceedingly small, and the assumption appears justified. For higher-order modes such as B′ of Fig. 17, it is clear that the field lines must have longitudinal components. This is illustrated by an edge view in Fig. 19. However, provided the width of a cell c is much greater than a half-wavelength, the longitudinal field intensity should be negligible compared to the transverse. Only for very high-order modes should this approximation begin to fail. Because the low-order modes of importance are essentially transverse electromagnetic, they are designated as TEM modes.

The plane mirror modes have a phase which is not constant over the mirror. This does not mean that it is impossible to space the mirrors for resonance of the entire field pattern. Actually, the phase delay for one transit is the same for every point on the wavefront. Therefore, if the plates are separated by the distance b plus an additional amount for the phase shift per transit of the mode desired, that mode should resonate in

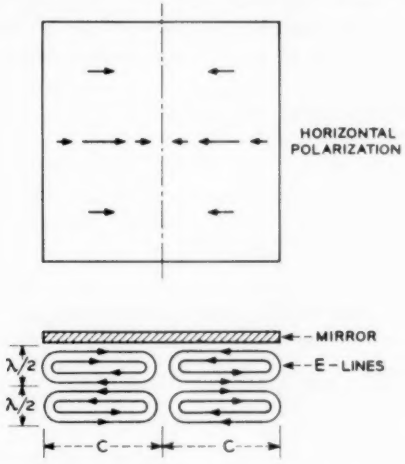


Fig. 19 — Field configuration of the TEM_{10} mode for square mirrors.

the interferometer. Other modes should not be resonant for this separation because they have different phase shifts per transit.

Since the field configurations of many of the normal modes of the interferometer are very similar to those of metal tube and parallel-plane waveguides, it is not surprising to find that simple waveguide theory can be used to predict certain characteristics of the interferometer modes. One of these characteristics is phase shift per transit. For instance, the field distributions of the normal modes for infinite strip mirrors are very similar to those of the TE modes of parallel-plane waveguide; also, by adding two orthogonally polarized TEM_{10} modes for circular plane mirrors, one obtains a field configuration which is very similar to that of the circular electric (TE_{01}) mode of circular waveguide (Fig. 20). Thus the amount of phase shift per transit computed for these modes of the interferometer agrees well with the phase shifts obtained for TE modes of parallel-plane waveguide and TE_{01} mode of circular waveguide. This is illustrated in Fig. 21. We see that agreement becomes better for larger values of N. This is because the similarity between field configurations becomes closer for larger values of N.

If we regard a uniform plane wave as being resolvable into a set of normal modes, there can be no such thing as a resonance for a uniform plane wave. Why then does it appear that there is such a resonance in passive optical interferometers? It is because for the usual optical case $a^2/b\lambda$ is in the thousands. The phase shifts per transit are extremely small, hence the mode resonances lie very close together in frequency. At the same time, the reflection coefficients of the best optical mirrors are so poor, and the Q of the interferometer is so low, that the resonance

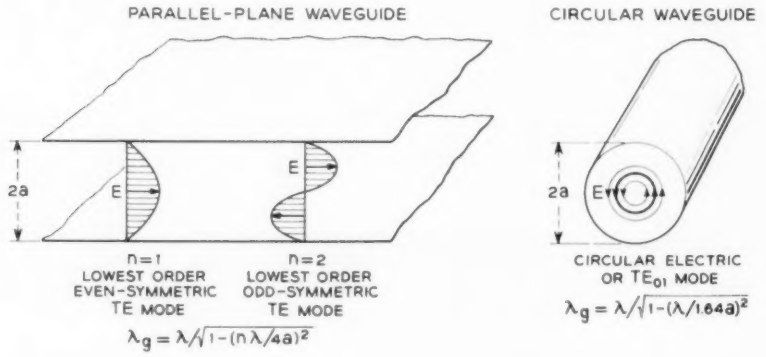


Fig. 20 — TE modes in a parallel-plane waveguide and circular electric mode in a circular waveguide.

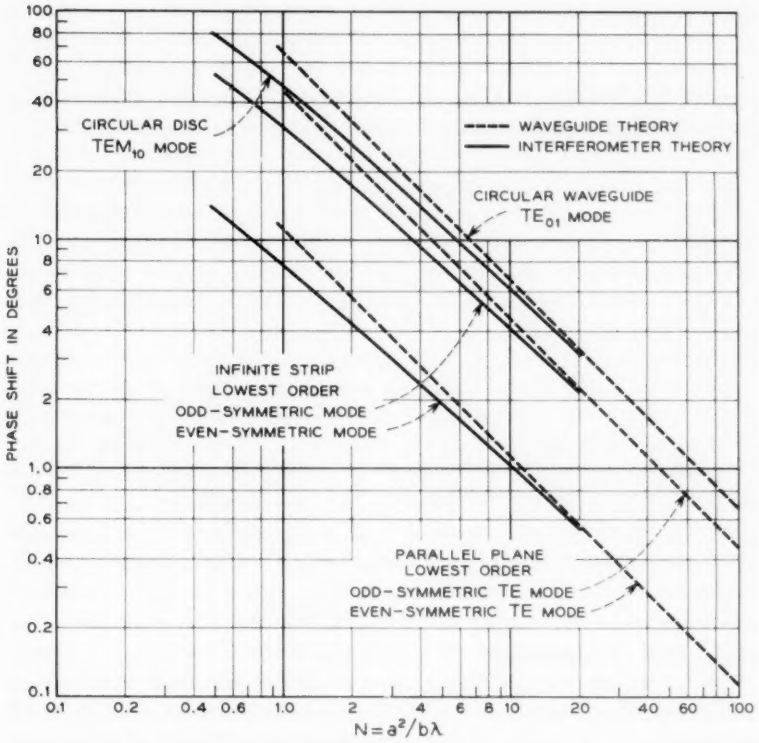


Fig. 21 — Comparison of computed phase shifts based on waveguide theory and on interferometer theory.

line width contains hundreds of normal mode resonances. Thus the uniform plane wave undergoes very little decomposition when resonated. Nevertheless, in the case of an active interferometer, the decomposition may be complete.

We now make use of the formula for the Q of a resonant waveguide cavity to compute the Q of an interferometer system. The Q of a resonant waveguide cavity is given by

$$Q = \frac{|R_1R_2e^{-2\alpha b}|}{1 - |R_1R_2e^{-2\alpha b}|}\left(\frac{2\pi b}{\lambda_g}\right)\left(\frac{\lambda_g}{\lambda}\right)^2, \tag{14}$$

where α is the attenuation constant of the waveguide and λ_g is the guide wavelength. For the interferometer we assume that α is zero and that λ_g is equal to λ, the free-space wavelength. The voltage reflection coeffi-

cients R_1 and R_2 for the two reflectors are given by

$$|R_1| = |R_2| = \sqrt{1 - \delta_r - \delta_d}, \tag{15}$$

where δ_r is the power loss in reflection and δ_d is the power loss due to spill-over. When these losses are very small, Q reduces to

$$Q \cong \frac{1}{\delta_r + \delta_d} 2\pi \frac{b}{\lambda}. \tag{16}$$

Hence

$$\frac{1}{Q} = \frac{1}{Q_r} + \frac{1}{Q_d}, \tag{17}$$

where

$$Q_r = \frac{2\pi b}{\lambda \delta_r}, \qquad Q_d = \frac{2\pi b}{\lambda \delta_d}. \tag{18}$$

The resonance line width at half-power points given as the change in electrical length of the resonator, $\Delta\varphi$, is

$$\begin{aligned} \Delta\varphi &= 2\pi \left(\frac{b}{\lambda}\right)\left(\frac{\Delta\lambda}{\lambda}\right) \\ &= \delta_r + \delta_d \text{ radians}, \end{aligned} \tag{19}$$

where we have substituted $(1/Q)$ for $(\Delta\lambda/\lambda)$.

Let us consider an interferometer having circular plane mirrors with $2a = 1$ cm, $b = 20$ cm, $\lambda = 5 \times 10^{-5}$ cm and a reflection loss of $\delta_r = 0.02$. In this case $N = a^2/b\lambda = 250$. Extrapolating the loss and phase shift curves of Figs. 8 and 9, we obtain diffraction loss $\delta_d = 9 \times 10^{-5}$ and phase shift for the dominant (TEM_{00}) mode $\varphi_d = 0.11$ degree. The diffraction loss is thus negligible compared to reflection loss, which limits the Q to a value of 1.25×10^8. The phase shift for the next higher order (TEM_{10}) mode is 0.30 degree and therefore it is separated from the dominant (TEM_{00}) mode by 0.19 degree or 0.0033 radian. The resonance line width, as given by (19), is 0.02 radian. Thus we see that TEM_{00} and TEM_{10} modes are not resolved. As the mirror separation is reduced or mirror size increased, more and more normal modes will become unresolved and a uniform plane wave will suffer less decomposition when resonated.

When an interferometer is filled with an active medium, the medium can compensate for the mirror losses and yield an enormously increased Q. Under these circumstances, the modes may be clearly resolved, and their Q's will be determined by the diffraction losses. If the gain of the

medium is increased until it compensates for mirror losses plus the diffraction loss of the lowest order mode, that mode will become unstable and oscillation can result. All higher-order modes will be stable and have positive net loss. If the gain of the medium is further increased, then many modes may become unstable. In starting from a quiescent condition, spontaneous emission can initiate a large number of characteristic waves in the interferometer. These may then start to grow, but the dominant mode will always grow faster and should saturate first. At saturation the steady-state field distribution will be considerably altered. The relative field at the edges of the mirrors should increase, thereby increasing the relative power loss. This can be described as a coupling of power into other modes as a result of the nonlinearity of the medium. No attempt has yet been made to analyze this situation. The linear theory is at present of most interest because it allows the computation of the starting conditions for oscillation.

With the development of the normal mode picture of interferometer operation and the computation of the losses for these modes, we may now ask if there is an optimum geometry for a maser interferometer which will permit oscillation for the lowest possible gain in the medium. We know that the power gained from the medium can be increased by increasing length. For very great lengths corresponding to the far-field region ($N < 0.1$), the power gained from the medium increases more rapidly than transmission loss as length is increased, and there must always be some length beyond which oscillations can occur. However, these lengths are too great to be of practical interest. In the near-field region ($N > 1$), represented by the curves of Fig. 8, the diffraction loss increases more rapidly than the medium gain. Therefore, if the reflection loss is sufficiently small, an optimum length may exist which is most favorable for oscillation.

To be more specific, let us consider a circular plane mirror interferometer. From Fig. 8 we find that the loss for the dominant mode may be represented by the expression

$$\delta_d = 0.207 \left(\frac{b\lambda}{a^2}\right)^{1.4}. \tag{20}$$

In order to find the optimum value of b to give a maximum Q, (20) and (18) are substituted in (17) and the resulting equation is differentiated with respect to b. For the optimum b, the diffraction loss is 2.5 times the reflection loss, and not equal to it, as might be supposed. Moreover, this result is general and holds for all modes and all shapes of plane mirrors represented in Fig. 8, provided the optimum falls on the straight-line

portions of the loss curves. Since the power supplied by the medium is proportional to the stored energy in the interferometer, while the power loss of the passive interferometer is just ω/Q times the stored energy, oscillation is most likely to occur when Q is a maximum. Fig. 22 illustrates the way the interferometer dimensions affect Q. If a given mirror diameter is chosen (as represented by the dashed line A), there is clearly an optimum distance b which will produce a maximum Q (intersection of lines A and B). However, if the distance b is held constant, there is no optimum value for a. The larger a, the higher will be Q, although it will approach a limiting value beyond which there is nothing to be gained by further increase of a.

As an example, let us assume a case where

$$\lambda = 10^{-4} \text{ cm},$$

$$2a = \text{plate diameter} = 2 \text{ cm},$$

$$\delta_r = \text{power reflection loss} = 0.001.$$

The optimum proportions require that δ_d be 0.0025, and for this, b is 435 cm and the resulting Q is 7.8×10^9. The length of 435 cm is probably impractically large for a maser. If b is reduced to a more reasonable value of 50 cm, the Q will drop to 3.14×10^9, which is the limiting value due to reflection loss. (The value assumed here for δ_r is already much lower than can be obtained from evaporated metal films and would require the technique of multilayered dielectric films.) In order to oscillate, the active medium would have to have a power amplification factor in excess of 1.00002 per centimeter of path.

In the case of confocal paraboloidal mirrors of 2 cm diameter, the optimum length turns out to be 8900 cm. If the diameter is reduced to 0.5 cm, the optimum length is still 530 cm, and for these proportions Q is 3.1×10^{10}. It is clear that with confocal mirrors the diffraction losses are negligible for any reasonable proportions of the interferometer.

One question of importance is whether there is an optimum set of dimensions which will discriminate against unwanted modes. It has sometimes been suggested that by making the mirror diameter small relative to the mirror spacing, "slant rays" will be more rapidly lost from the system. However, from Fig. 8 it can be seen that the ratios of the losses for the several modes is independent of N provided N is greater than 1. Thus, if diffraction losses predominate, there is no way of discriminating against unwanted modes by juggling dimensions. The limiting amount of discrimination is merely governed by the ratio of the losses for the different modes, which is independent of the dimensions. However, if reflection losses predominate, the discrimination between

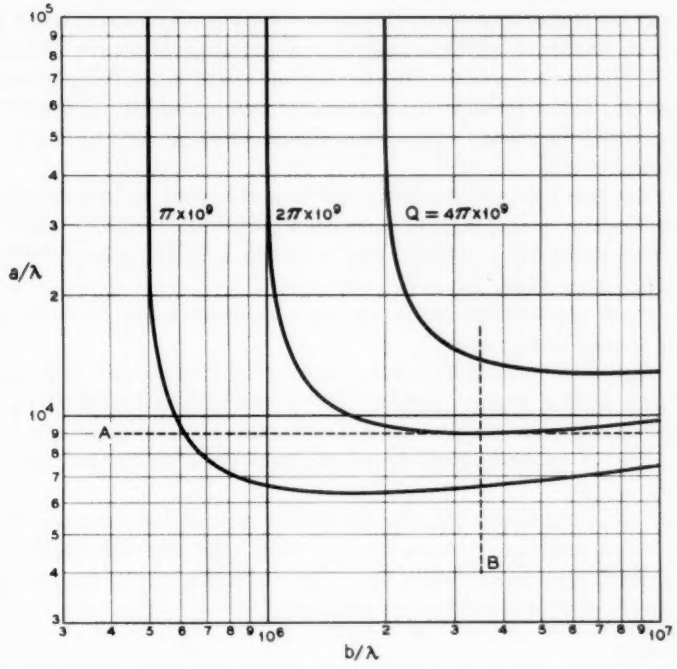


Fig. 22 — Interferometer dimensions for constant Q. (Circular plane mirrors, reflection loss = $\delta_r = 0.001$.)

lower-order modes would be almost nonexistent and it would be advantageous to increase mirror separation and/or decrease mirror dimensions so as to make diffraction losses predominate. In the case of the confocal mirrors, the loss ratios between modes are not constant (Fig. 15) although, for values of N larger than those shown, they may become so. At any rate, for values of N close to unity, a small amount of increased discrimination against higher order modes can be obtained by making the mirrors *larger*.

V. CONCLUSIONS

Diffraction studies carried out on the IBM computer have led to the following conclusions:

1. Fabry-Perot interferometers, whether of the plane or concave mirror type, are characterized by a discrete set of normal modes which can be defined on an iterative basis. The dominant mode has a field intensity which falls to low values at the edges of the mirrors, thereby

causing the power loss due to diffractive spillover to be much lower than would be predicted on the assumption of uniform plane wave excitation.

2. Uniform plane waves are not normal modes for a flat-plate interferometer. Consequently, interferometer resonances do not exist for "slant rays," i.e., plane waves traveling at an angle with respect to the longitudinal axis.

3. The losses for the dominant mode of the plane mirror system are so low that for most practical geometries performance will be limited by reflection losses and scattering due to aberrations. For confocal mirrors the diffraction losses are even lower.

4. There are no higher-order modes with losses lower than the dominant (lowest-order) mode.

5. The ratio of diffraction losses between the modes investigated for the plane mirror system is independent of the interferometer dimensions in the range of interest. Therefore, if diffraction losses predominate, there is no way of proportioning the interferometer so as to favor any one mode.

The computer technique we employed is general and versatile. It can be used for studying mirrors having rather arbitrary but small curvatures. With little modification, the same technique can be used to study the effects of aberration and misalignment.

APPENDIX A

Rectangular Plane Mirrors

The geometry for rectangular plane mirrors parallel to the xy plane is shown in Fig. 2. According to (1), the iterative equation for computing the field at the surface of mirrors is

$$u_{q+1}(x_2, y_2) = \frac{j}{2\lambda} \int_{-c}^{c} \int_{-a}^{a} u_q(x_1, y_1) \frac{e^{-jkR}}{R} \left(1 + \frac{b}{R}\right) dx_1 dy_1, \quad (21)$$

where

$$R = \sqrt{b^2 + (x_1 - x_2)^2 + (y_1 - y_2)^2}.$$

If b/a and b/c are large, (21) can be reduced to

$$u_{q+1}(x_2, y_2) = \frac{je^{-jkb}}{\lambda b} \int_{-c}^{c} \int_{-a}^{a} u_q(x_1, y_1) e^{-jk[(x_1-x_2)^2+(y_1-y_2)^2]/2b} dx_1 dy_1, \quad (22)$$

which is valid for $(a^2/b\lambda) \ll (b/a)^2$ and $(c^2/b\lambda) \ll (b/c)^2$.* The corresponding integral equation is

* Actually, the stringency of this requirement can be relaxed somewhat for lower-order modes in which field intensities near the edges of the mirror are rather low. We have made check computations for the case $a^2/b\lambda = 5$ and $(b/a)^2 = 25$ and have found that the results based on the exact equation and on the approximate equation are in essential agreement.

$$v(x_2, y_2) = \gamma \int_{-c}^{c} \int_{-a}^{a} K(x_2, x_1; y_2, y_1) v(x_1, y_1)\, dx_1 dy_1, \tag{23}$$

where

$$K(x_2, x_1; y_2, y_1) = \frac{j}{\lambda b} e^{-jk[(x_1-x_2)^2+(y_1-y_2)^2]/2b} \tag{23a}$$

and the factor e^{-jkb} is absorbed in γ.

Here, the kernel of the integral equation is separable in x and y. If the distribution function v is assumed to be of the form

$$v(x,y) = v_x(x) v_y(y) \tag{24}$$

it is possible to separate (23) into two equations, one involving x only and the other involving y only; that is,

$$v_x(x_2) = \gamma_x \int_{-a}^{a} K_x(x_2, x_1) v_x(x_1)\, dx_1, \tag{25a}$$

$$v_y(y_2) = \gamma_y \int_{-c}^{c} K_y(y_2, y_1) v_y(y_1)\, dy_1, \tag{25b}$$

with

$$K_x = \frac{e^{j(\pi/4)}}{\sqrt{\lambda b}} e^{-jk(x_1-x_2)^2/2b}, \tag{25c}$$

and

$$K_y = \frac{e^{j(\pi/4)}}{\sqrt{\lambda b}} e^{-jk(y_1-y_2)^2/2b}. \tag{25d}$$

The product of the eigenvalues γ_x and γ_y is equal to the eigenvalue γ in (23).

It remains to be shown that (25a) through (25d) represent integral equations for infinite strip mirrors. Let us consider a pair of infinite strip mirrors of width $2a$ and separated by b. The iterative equation for computing the field at the mirrors can be derived from (1). It is

$$u_{q+1}(x_2) = \frac{e^{j(\pi/4)}}{2\sqrt{\lambda}} \int_{-a}^{a} u_q(x_1) \frac{e^{-jk\rho}}{\sqrt{\rho}} \left(1 + \frac{b}{\rho}\right) dx_1, \tag{26}$$

where

$$\rho = \sqrt{b^2 + (x_1 - x_2)^2}.$$

For $(a^2/b\lambda) \ll (b/a)^2$, (26) reduces to

$$u_{q+1}(x_2) = \frac{e^{j[(\pi/4)-kb]}}{\sqrt{\lambda b}} \int_{-a}^{a} u_q(x_1)\, e^{-jk(x_1-x_2)^2/2b}\, dx_1. \tag{27}$$

The corresponding integral equation is

$$v(x_2) = \gamma \int_{-a}^{a} K(x_2, x_1)\, v(x_1)\, dx_1, \tag{28}$$

where

$$K(x_2, x_1) = \frac{e^{j(\pi/4)}}{\sqrt{\lambda b}}\, e^{-jk(x_1 - x_2)^2/2b} \tag{28a}$$

and the factor e^{-jkb} is absorbed in γ. We see that (25) and (28) are identical in form.

APPENDIX B

Circular Plane Mirrors

Assuming approximately plane waves propagating normally to the circular plane mirrors (Fig. 3), the iterative equation for computing the steady-state field distribution can be written as

$$u_{q+1}(r_2, \varphi_2) = \frac{j}{2\lambda} \int_0^a \int_0^{2\pi} u_q(r_1, \varphi_1)\, \frac{e^{-jkR}}{R} \left(1 + \frac{b}{R}\right) r_1\, d\varphi_1\, dr_1, \tag{29}$$

where

$$R = \sqrt{b^2 + r_1^2 + r_2^2 - 2r_1 r_2 \cos(\varphi_1 - \varphi_2)}.$$

If b/a is large, (29) simplifies to

$$u_{q+1}(r_2, \varphi_2) = \frac{je^{-jkb}}{\lambda b} \int_0^a \int_0^{2\pi} u_q(r_1, \varphi_1) \cdot e^{-jk[(r_1^2 + r_2^2)/2b - (r_1 r_2/b)\cos(\varphi_1 - \varphi_2)]} r_1\, d\varphi_1\, dr_1, \tag{30}$$

which is valid for $(a^2/b\lambda) \ll (b/a)^2$.*

The integral equation corresponding to (30) is

$$v(r_2, \varphi_2) = \gamma \int_0^a \int_0^{2\pi} K(r_2, \varphi_2; r_1, \varphi_1) v(r_1, \varphi_1) r_1\, d\varphi_1\, dr_1, \tag{31}$$

with

$$K(r_2, \varphi_2; r_1, \varphi_1) = \frac{j}{\lambda b}\, e^{-jk[(r_1^2 + r_2^2)/2b - (r_1 r_2/b)\cos(\varphi_1 - \varphi_2)]} \tag{31a}$$

* Comments in Appendix A regarding the stringency of this requirement are also applicable herein.

and where the factor e^{-jkb} is absorbed in γ. Making use of the relation[10]

$$e^{jn[(\pi/2)-\varphi_2]}J_n\left(k\frac{r_1r_2}{b}\right) = \frac{1}{2\pi}\int_0^{2\pi} e^{jk(r_1r_2/b)\cos(\varphi_1-\varphi_2)-jn\varphi_1}\,d\varphi_1 \qquad (32)$$

and integrating (31) with respect to φ_1, it is seen that

$$v(r,\varphi) = R_n(r)e^{-jn\varphi}, \qquad (n = \text{integer}) \qquad (33)$$

satisfies (31). The function $R_n(r)$ satisfies the reduced integral equation

$$R_n(r_2)\sqrt{r_2} = \gamma_n\int_0^a K_n(r_2, r_1)R_n(r_1)\sqrt{r_1}\,dr_1, \qquad (34)$$

with

$$K_n(r_2, r_1) = \frac{j^{n+1}k}{b}J_n\left(k\frac{r_1r_2}{b}\right)\sqrt{r_1r_2}\,e^{-jk(r_1^2+r_2^2)/2b}, \qquad (34a)$$

where J_n is a Bessel function of the first kind and nth order.

APPENDIX C

Confocal Spherical or Paraboloidal Mirrors

For confocal spherical mirrors of circular cross section (Fig. 4), the iterative equation corresponding to (29) is

$$u_{q+1}(r_2, \varphi_2) = \frac{j}{2\lambda}\int_0^a\int_0^{2\pi} u_q(r_1, \varphi_1)\frac{e^{-jkR}}{R}\left(1+\frac{b_1}{R}\right)r_1\,d\varphi_1\,dr_1 \qquad (35)$$

where

$$R = \sqrt{b_1^2 + r_1^2 + r_2^2 - 2r_1r_2\cos(\varphi_1 - \varphi_2)}.$$

The distance b_1 is given by

$$b_1 = b - \Delta_1 - \Delta_2 \qquad (36)$$

where, for confocal spherical mirrors,

$$\Delta_i = b - \sqrt{b^2 - r_i^2} \qquad i = 1,2. \qquad (36a)$$

If b/a is large, the distance Δ_i is given approximately by

$$\Delta_i \cong r_i^2/2b \qquad i = 1,2, \qquad (37)$$

which is exact for confocal paraboloids. In this case (35) simplifies to

$$u_{q+1}(r_2, \varphi_2) = \frac{je^{-jkb}}{\lambda b} \int_0^a \int_0^{2\pi} u_q(r_1, \varphi_1)\, e^{jk(r_1 r_2/b)\cos(\varphi_1 - \varphi_2)} r_1\, d\varphi_1\, dr_1, \quad (38)$$

which is valid for $(a^2/b\lambda) \ll (b/a)^2$.

The integral equation corresponding to (38) is

$$v(r_2, \varphi_2) = \gamma \int_0^a \int_0^{2\pi} K(r_2, \varphi_2; r_1, \varphi_1) v(r_1, \varphi_1) r_1\, d\varphi_1\, dr_1, \quad (39)$$

with

$$K(r_2, \varphi_2; r_1, \varphi_1) = \frac{j}{\lambda b}\, e^{jk(r_1 r_2/b)\cos(\varphi_1 - \varphi_2)} \quad (40)$$

and where the factor e^{-jkb} is absorbed in γ. Just as in the case of circular plane mirrors, it can be shown that

$$v(r,\varphi) = S_n(r) e^{-jn\varphi} \qquad (n = \text{integer}) \quad (41)$$

satisfies (39). The function $S_n(r)$ satisfies the reduced integral equation

$$S_n(r_2)\sqrt{r_2} = \gamma_n \int_0^a K_n(r_2, r_1) S_n(r_1) \sqrt{r_1}\, dr_1, \quad (42)$$

with

$$K_n(r_2, r_1) = \frac{j^{n+1}k}{b} J_n\left(k\frac{r_1 r_2}{b}\right) \sqrt{r_1 r_2}. \quad (42a)$$

REFERENCES

1. Schawlow, A. L. and Townes, C. H., Infrared and Optical Masers, Phys. Rev., **112,** 1958, p. 1940.
2. Maiman, T. H., Stimulated Optical Radiation in Ruby, Nature, **187,** 1960, p. 493.
3. Collins, R. J., Nelson, D. F., Schawlow, A. L., Bond, W., Garrett, C. G. B. and Kaiser, W., Coherence, Narrowing, Directionality and Relaxation Oscillations in the Light Emission from Ruby, Phys. Rev. Letters, **5,** 1960, p. 303.
4. Silver, S., *Microwave Antenna Theory and Design*, McGraw-Hill, New York, 1949, p. 167.
5. Hildebrand, F. B., *Methods of Applied Mathematics*, Prentice-Hall, Englewood Cliffs, N. J., 1952.
6. Connes, P., Increase of the Product of Luminosity and Resolving Power of Interferometers by Using a Path Difference Independent of the Angle of Incidence, Revue d'Optique, **35,** 1956, p. 37.
7. Lovitt, W. V., *Linear Integral Equations*, Dover Publications, New York, 1950, pp. 129; 137.
8. Goubau, G., and Schwering, F., On the Guided Propagation of Electromagnetic Wave Beams, URSI-IRE Spring Meeting, May 1960, Washington, D. C.
9. Boyd, G. D., and Gordon, J. P., Confocal Multimode Resonator for Millimeter through Optical Wavelength Masers, this issue, p. 489.
10. Stratton, J. A., *Electromagnetic Theory*, McGraw-Hill, New York, 1941, p. 372.

Confocal Multimode Resonator for Millimeter Through Optical Wavelength Masers

By G. D. BOYD and J. P. GORDON

(Manuscript received September 12, 1960)

Multimode resonators of high quality factor will very likely play a significant role in the development of devices, such as the maser, which operate in the millimeter through optical wavelength range. It has been suggested that a plane-parallel Fabry-Perot interferometer could act as a suitable resonator. In this paper a resonator consisting of two identical concave spherical reflectors, separated by any distance up to twice their common radius of curvature, is considered.

Mode patterns and diffraction losses for the low-loss modes of such a resonator are obtained analytically, using an approximate method which was suggested by W. D. Lewis. The results show that the diffraction losses are generally considerably lower for the curved surfaces than for the plane surfaces. Diffraction losses and mode volume are a minimum when the reflector spacing equals the common radius of curvature of the reflectors. For this case the resonator may be termed confocal. A further property of the concave spherical resonator is that the optical alignment is not extremely critical.

I. INTRODUCTION

Schawlow and Townes[1] proposed that coherent amplification could be achieved in the infrared through optical regions of the frequency spectrum by maser techniques. At such frequencies multimode resonators are necessary to achieve reasonable dimensions and high Q. They and Prokhorov[2] and Dicke[3] have suggested as a resonator two plane-parallel reflecting planes, known as a Fabry-Perot interferometer, or etalon.[4]

In Fabry-Perot resonators the major factors contributing to the Q (i.e., resolving power) are reflection losses and diffraction losses. Reflection losses result from absorption in the reflectors, and from transmission

through them. At optical frequencies a very good layered dielectric reflector[5] can have a 99½ per cent reflection coefficient. Diffraction losses result from the finite aperture of the reflectors and from imperfections in their "flatness."

Fox and Li have shown in the accompanying paper[6] that modes, in the sense of a self-reproducing field pattern, exist for an open structure such as a Fabry-Perot interferometer. They also have recognized that the diffraction losses of a plane-parallel Fabry-Perot are very much less than those obtained by assuming a uniform intensity distribution over the reflector and the Fraunhofer far field diffraction angle. They have made numerical self-consistent field calculations based on Huygens' principle to determine the actual diffraction losses and mode patterns.

In interferometry using a Fabry-Perot resonator, one normally excites a system of plane waves traveling at certain discrete angles to the axis. Constructive interference at each of these discrete angles, as is appropriate to ring order, wavelength and spacing, results in a pattern of concentric bright rings. Schawlow and Townes indicated that each ring of the interference pattern is *not* a pure mode of the resonator but an infinite sum of such modes, each representing a different field pattern over the reflector. This idea has been given much substance by the work of Fox and Li.

The plane-parallel Fabry-Perot is not necessarily ideal, however, as a high-frequency multimode resonator. A resonator formed by two spherical reflectors of equal curvature separated by their common radius of curvature is considered in detail in this paper. The focal length of a spherical mirror is one-half of its radius of curvature. Therefore the focal points of the reflectors are coincident and the resonator is termed confocal. G. W. Series, Fox and Li[6] and Lewis[7] have also suggested the confocal resonator. Lewis has recognized that it would have lower diffraction losses than the plane-parallel Fabry-Perot and has described the analytic solution presented here.

The use of confocal reflectors as an interferometer has been described by Connes.[8] The adjustment of the spherical Connes interferometer is trivial compared to the Fabry-Perot. Parallelism between the reflectors is not a strict requirement, the only fine adjustment therefore being the spacing between the surfaces. Parabolic surfaces may also be used, but they have an axis and thus lose the advantage of ease of adjustment.

II. RESONATOR QUALITY FACTOR

Resonator quality factor, or Q, is defined as

$$Q = \omega \frac{\text{energy stored}}{\text{energy lost per second}}. \tag{1}$$

Consider an interferometer consisting of two reflecting surfaces separated by a distance d which is large compared to the wavelength in the medium λ. By considering waves bouncing back and forth between the surfaces, one may derive an approximate Q as

$$Q = \frac{2\pi d}{\alpha\lambda}, \tag{2}$$

where α is the fractional power loss per bounce from a reflector and is the sum of diffraction and reflection losses. This is to be compared to the resolving power derived in optics[9] as

$$R = \frac{2\pi d\sqrt{r}}{\lambda(1 - r)}, \tag{3}$$

where the power reflection coefficient per bounce is $r \equiv 1 - \alpha$. Resolving power is thus synonymous with Q within the small loss approximation of (2).

If diffraction losses are small compared with reflection losses, then resonator Q is proportional to the spacing between the reflecting surfaces. For a given reflector aperture size, the resonator Q will continue to increase with the spacing d between the reflectors until the diffraction losses become roughly comparable with the reflection losses. Further increase in spacing then decreases the Q because of increasing diffraction losses.

III. MODES AND DIFFRACTION LOSSES OF A CONFOCAL RESONATOR

All resonator dimensions are assumed large compared to a wavelength; the modes and diffraction losses of the confocal resonator are therefore obtainable from a self-consistent field analysis using Huygens' principle.[10] A confocal resonator is considered, with identical spherical reflectors of radius b, as shown in Fig. 1. Assume the field to be linearly polarized over the P' surface in the y direction and given by $E_0 f_m(x')g_n(y')$, where E_0 is a constant amplitude factor and $f_m(x')$ and $g_n(y')$ are the field variations over the aperture. At point $P(x,y)$ on the other surface, one computes the electric field by summing over contributions from the differential Huygens sources at all points $P'(x',y')$. The result is

$$E_y = \int_{S'} \frac{ik(1 + \cos\theta)}{4\pi\rho} e^{-ik\rho} E_0 f_m(x')g_n(y')\, dS'. \tag{4}$$

Here ρ is the distance between P and P', θ is the angle between the line $P'P$ and the normal to the reflector surface at P', and k is the propagation constant of the medium between the reflectors. Note that $k =$

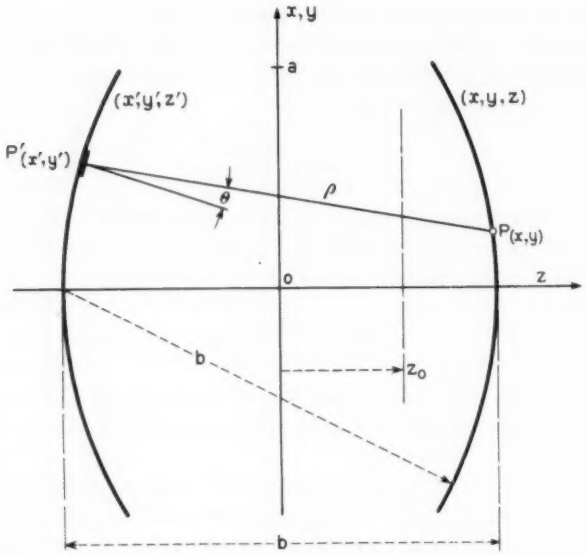


Fig. 1 — Confocal resonator with spherical reflectors.

$2\pi/\lambda$, where λ is the wavelength in the medium. The electric field in the xz plane is approximately zero. The reflector is assumed *square* and of dimension $2a$, which is small compared to the spacing b (since the confocal spacing is under consideration $d = b$), and thus θ is very nearly zero. The medium is assumed to fill all space.

The normal modes or eigenfunctions of the confocal resonator are obtained by requiring that the field distribution over $x'y'$ reproduce itself within a constant over the xy aperture, and thus $E_y = E_1 f_m(x) g_n(y)$, where $E_1 = \sigma_m \sigma_n E_0$. The proportionality factor $\sigma_m \sigma_n$ is generally complex, giving both amplitude and phase changes. The resulting integral equation is

$$\sigma_m \sigma_n f_m(x) g_n(y) = \iint_{-a}^{+a} \frac{ik}{2\pi\rho} e^{-ik\rho} f_m(x') g_n(y')\, dx' dy' . \tag{5}$$

The distance ρ varies only a small amount for small apertures and thus may be replaced by the separation b except in the exponential phase term. For x and y small compared to b one can show that

$$\frac{\rho}{b} = 1 - \frac{xx' + yy'}{b^2} + \frac{w^2 w'^2}{4b^4} + \cdots , \tag{6}$$

where $w^2 = x^2 + y^2$. The third term makes a negligible contribution to the phase when $a^2/b\lambda \ll b^2/a^2$. Note that in this approximation one cannot distinguish between spherical and parabolic surfaces. In terms of some dimensionless variables

$$c \equiv \frac{a^2 k}{b} = 2\pi\left(\frac{a^2}{b\lambda}\right) \qquad X \equiv \frac{x\sqrt{c}}{a}, \qquad Y \equiv \frac{y\sqrt{c}}{a}, \tag{7}$$

and with $F_m(X) \equiv f_m(x)$ etc., (5) becomes

$$\sigma_m\sigma_n F_m(X)G_n(Y) = \frac{ie^{-ikb}}{2\pi}\int_{-\sqrt{c}}^{+\sqrt{c}} F_m(X')e^{+iXX'}\,dX' \cdot \int_{-\sqrt{c}}^{+\sqrt{c}} G_n(Y')e^{+iYY'}\,dY'. \tag{8}$$

Slepian and Pollak[11] have considered the following integral equation:

$$F_m(X) = \frac{1}{\sqrt{2\pi\chi_m}}\int_{-\sqrt{c}}^{+\sqrt{c}} F_m(X')e^{+iXX'}\,dX'. \tag{9}$$

This is a homogeneous Fredholm equation of the second kind with $e^{iXX'}$ as the kernel. It is often referred to as a finite Fourier transform. They have shown solutions to be

$$F_m(c,\eta) \propto S_{0m}(c,\eta), \tag{10}$$

$$\chi_m = \sqrt{\frac{2c}{\pi}}\, i^m R_{0m}^{(1)}(c,1), \qquad m = 0, 1, 2, \cdots, \tag{11}$$

where $S_{0m}(c,\eta)$ and $R_{0m}^{(1)}(c,1)$ are respectively the angular and radial wave functions in prolate spheroidal coordinates as defined by Flammer,[12] and where $\eta = X/\sqrt{c} = x/a$ and $\eta = Y/\sqrt{c} = y/a$ respectively for $F_m(X)$ and $G_n(Y)$. There is an infinite number of eigenfunctions and corresponding eigenvalue solutions to (9) for any value of c. Flammer[12] gives values of these functions for $c \leqq 5$ and Slepian and Pollak[11,13] have computed the eigenvalues χ_m for the important region of $c > 5$.

The eigenfunction solutions of (8) are thus the spheroidal wave functions $S_{0m}(c,x/a)S_{0n}(c,y/a)$. The eigenfunctions are real; therefore, the reflecting surfaces are of constant phase. The eigenvalues are

$$\sigma_m\sigma_n = \chi_m\chi_n i\, e^{-ikb}. \tag{12}$$

The phase shift between the two reflecting confocal surfaces equals the phase angle of $\sigma_m\sigma_n$. For resonance the round-trip phase shift must

equal an integer q times 2π. From (11) and (12), one finds therefore

$$2\pi q = 2\left|\frac{\pi}{2} - kb + (m+n)\frac{\pi}{2}\right|. \tag{13}$$

Since $k = 2\pi/\lambda$, one obtains for the condition of resonance

$$\frac{4b}{\lambda} = 2q + (1 + m + n). \tag{14}$$

The confocal resonator is seen to have resonances only for integer values of the quantity $4b/\lambda$. If $4b/\lambda$ is odd, $(m + n)$ must be even, likewise if $4b/\lambda$ is even, $(m + n)$ must be odd. Note that considerable degeneracy exists in the spectrum; increasing $(m + n)$ by two and decreasing q by unity gives the same frequency. The degenerate modes are orthogonal over the reflector surface since they satisfy the integral (5) with different eigenvalues. The modes have negligible axial electric and magnetic fields and thus will be designated by TEM_{mnq}, where m and n equal 0, 1, 2, $\cdots$, and refer to variations in the x and y directions, while q equals the number of half-guide wavelength variations in the z direction between reflectors.

The fractional energy loss per reflection due to diffraction effects is given by

$$\alpha_D = 1 - |\sigma_m\sigma_n|^2 = 1 - |\chi_m\chi_n|^2. \tag{15}$$

The function $1 - |\chi_m|^2$ versus c is shown in Fig. 2 for $m = 0, 1, 2$. It can be shown that Fig. 2 also gives the diffraction losses for an infinite cylindrical reflector strip of width $2a$ and radius of curvature b. The diffraction losses for various TEM_{mnq} modes are shown in Fig. 3. Note that TEM_{uvq} and TEM_{vuq} $(u \neq v)$ have the same diffraction losses; also that the diffraction losses of the TEM_{02q} and TEM_{12q} are so nearly equal that they can be plotted as one curve. As indicated previously, these last two types of modes cannot both be resonant at the same frequency. Note that the losses are primarily determined by the higher of the transverse mode numbers m, n, regardless of the field polarization.

In Fig. 3 the results of Fox and Li[6] for the plane-parallel resonator with circular reflectors are also shown. The diffraction losses for the confocal resonator are seen to be orders of magnitude *smaller* than for the plane parallel resonator. Fox and Li have also obtained numerical results for the confocal resonator with circular cross section of radius a. These are in good agreement with the results presented here, allowing for the fact that in this paper the reflectors have a square cross section of width $2a$.

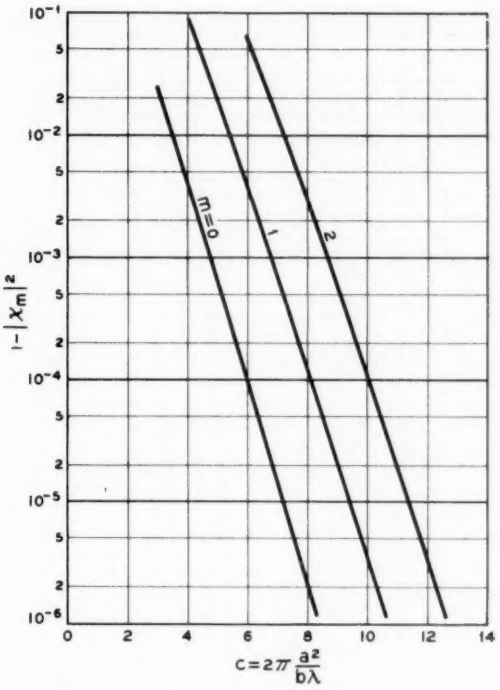


Fig. 2 — Eigenvalues of integral equation; also the diffraction losses of an infinitely long cylindrical reflector of width $2a$.

If one approximates the diffraction loss curve by a function $\alpha_D = A \times 10^{-B(a^2/b\lambda)}$, one may then show for a given reflection loss and reflector radius a that the resonator Q is a maximum as a function of the confocal spacing b when the reflection loss equals $[2.30B(a^2/b\lambda) - 1]$ times the diffraction loss. For the TEM_{00q} mode, $A = 10.9$ and $B = 4.94$; thus, if $a^2/b\lambda = 0.8$, then the diffraction loss is approximately one-eighth of the reflection loss.

The diffraction loss for the plane-parallel case assuming a uniform field and phase distribution and a diffraction angle of $\theta = \lambda/2a$ is also shown. This diffraction angle corresponds to the first Fraunhofer minimum in far field theory. For a square (or circular) reflector of side $2a$ the diffraction loss is approximately

$$\alpha_D \approx \left(\frac{a^2}{b\lambda}\right)^{-1}. \tag{16}$$

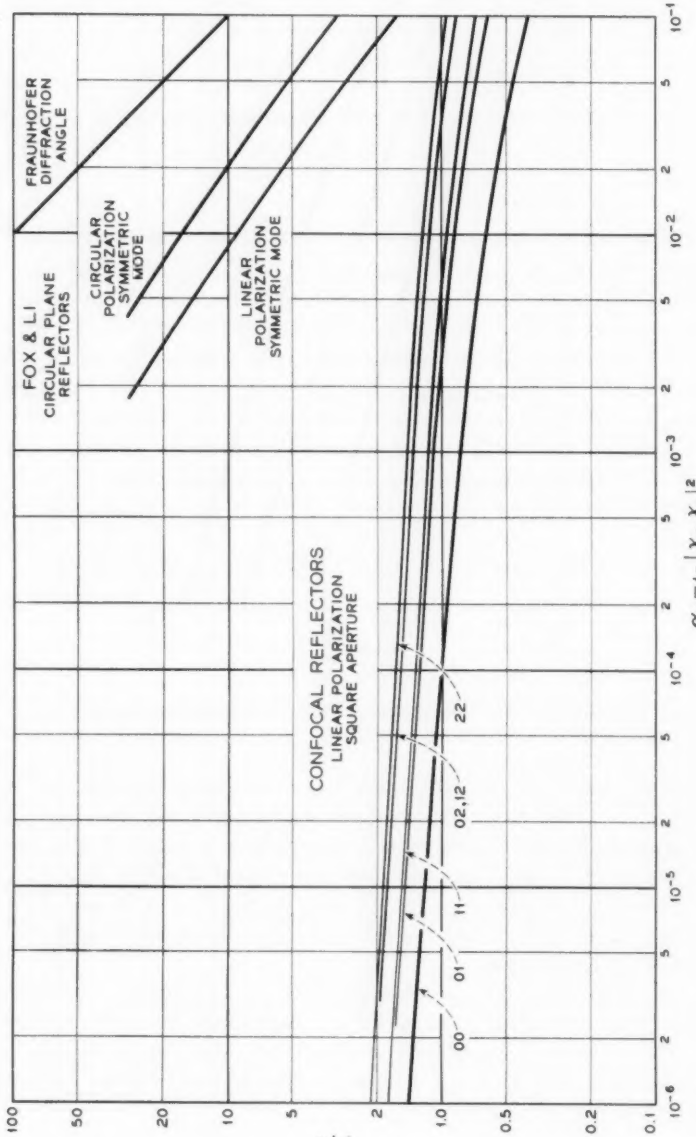


Fig. 3 — Diffraction losses for confocal and plane-parallel[6] resonators.

Fig. 3 clearly demonstrates the inadequacy of the assumption of uniform intensity distribution.

Though the eigenvalues given by (12) must be known accurately the eigenfunctions are only of approximate interest. Flammer[12] shows that, in the approximation of $\eta^2 \ll 1$ (near the center of the reflector), (10) becomes

$$F_m(X) \approx \frac{\Gamma\left(\frac{m}{2}+1\right)}{\Gamma(m+1)} H_m(X) e^{-\frac{1}{2}X^2} = \frac{\Gamma\left(\frac{m}{2}+1\right)}{\Gamma(m+1)} (-1)^m e^{+\frac{1}{2}X^2} \frac{d^m}{dX^m} e^{-X^2}. \tag{17}$$

The mode shape is thus approximately a Gaussian times a Hermite polynomial $H_m(X)$. The gamma function is arbitrarily chosen as normalization such that $F_m(X = 0) = \pm 1$ for m even:

$$\begin{aligned} F_0(c,\eta) &= e^{-\frac{1}{2}c\eta^2}, \\ F_1(c,\eta) &= \sqrt{\pi c}\,\eta\, e^{-\frac{1}{2}c\eta^2}, \\ F_2(c,\eta) &= (2c\eta^2 - 1)e^{-\frac{1}{2}c\eta^2}. \end{aligned} \tag{18}$$

The approximation involved in (17) fails away from the center of the reflector. For reasonably large values of c, however, the field is weak there, and of little interest. The diffraction losses were previously obtained from (15). Curves representing (18) for various values of c are shown in Fig. 4. The dotted curves for $c = 5$ are the true eigenfunctions $S_{0m}(c,\eta)$ as obtained from Flammer.[12]

The exponential dependence of the electric field on $c\eta^2$, which is independent of the reflector half-width a, leads one to define a "spot size" at the reflector of radius $w = w_s$, where $w^2 = x^2 + y^2$, at which the exponential term falls to e^{-1}:

$$w_s = \sqrt{\frac{b\lambda}{\pi}}. \tag{19}$$

The only effect of increasing the reflector width $2a$ is to reduce the diffraction losses; the spot size is unaffected.

If one allows the reflectors to be somewhat lossy or partially transparent, then the resonator Q is reduced over that implied by diffraction losses alone. The field distribution, i.e., the mode pattern, is *not* seriously affected so long as the losses are small and fairly uniform over the plates.

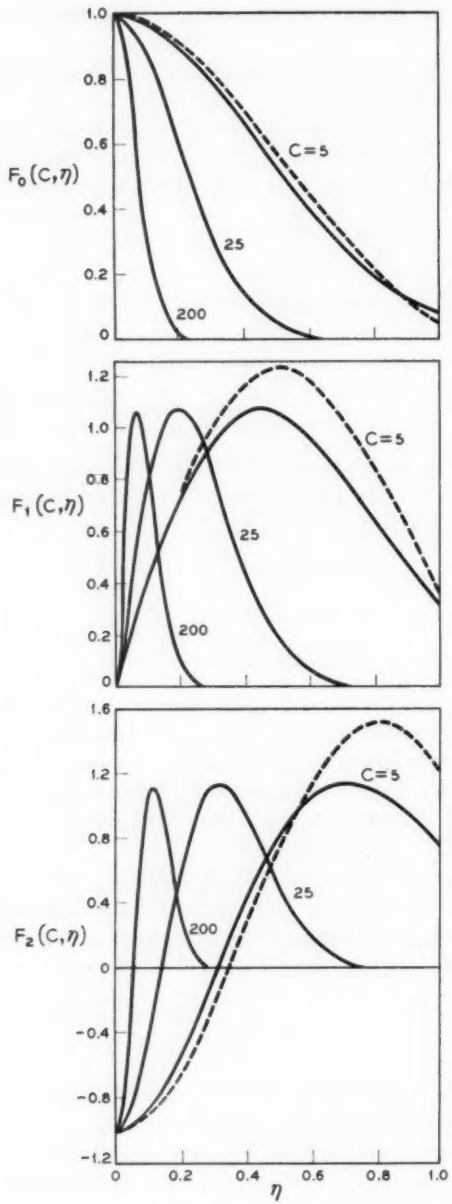


Fig. 4 — Approximate field amplitude variation versus normalized radius for various modes. The exact dependence given by the angular prolate spheroidal function $S_{0m}(c, \eta)$ is shown by dashed lines.

The electric field patterns derived thus far have all been linearly polarized. Fox and Li[6] have recognized that, by superimposing the TEM_{01q} mode linearly polarized in the x direction and the TEM_{10q} mode linearly polarized in the y direction, the lowest-order circular electric mode can result, and it has the same diffraction losses as the linearly polarized TEM_{01q} mode. Many other polarization configurations can be obtained in this manner.

IV. FIELDS OF THE CONFOCAL RESONATOR

The field over the confocal aperture has been obtained in the preceding section. The field over an arbitrary plane $z = z_0$, as in Fig. 1, is also obtainable by a straightforward application of Huygens' principle as stated in (4). The arbitrary plane z_0 may be placed outside the confocal geometry as well as inside provided one takes into account the transmission loss of the reflector. The field distribution over the confocal surface is given by $F_m(c,x/a)G_n(c,y/a)$. For large c the spheroidal functions may be approximated by the Gaussian-Hermite functions. The integral can be evaluated in the limit of $c \to \infty$.

Within these approximations, the traveling wave field of the confocal resonator resulting from the field at one of the reflectors is given by

$$\frac{E(x,y,z_0)}{E_0} = \sqrt{\frac{2}{1+\xi^2}}\,\frac{\Gamma\left(\frac{m}{2}+1\right)\Gamma\left(\frac{n}{2}+1\right)}{\Gamma(m+1)\Gamma(n+1)}\,H_m\left(X\sqrt{\frac{2}{1+\xi^2}}\right)$$
$$\cdot H_n\left(Y\sqrt{\frac{2}{1+\xi^2}}\right)\exp\left[-\frac{kw^2}{b(1+\xi^2)}\right] \tag{20}$$
$$\cdot\exp\left(-i\left\{k\left[\frac{b}{2}(1+\xi)+\frac{\xi}{1+\xi^2}\frac{w^2}{b}\right]-(1+m+n)\left(\frac{\pi}{2}-\varphi\right)\right\}\right),$$

where

$$w^2 = x^2 + y^2,$$
$$\xi = \frac{2z_0}{b}, \tag{21}$$
$$\tan\varphi = \frac{1-\xi}{1+\xi}.$$

When the reflecting surface is made partially transparent, as will be the case with optical or infrared masers, the field of the transmitted wave will be a traveling wave as given in (20) reduced by the transmission coefficient of the reflector. Within the resonator, the field will be a

standing wave. The transverse standing wave is as given in (20) except that the exponential phase function is replaced by the sine function.

The surface of constant phase which intersects the axis at z_0 as obtained from (20) is given approximately by

$$z - z_0 \approx - \frac{\xi}{1 + \xi^2} \frac{w^2}{b}, \tag{22}$$

neglecting the small variation in φ due to variation in z. This surface is spherical, within the approximations of this paper, and has a radius of curvature b' given by

$$b' = \left| \frac{1 + \xi^2}{2\xi} \right| b. \tag{23}$$

At $\xi = \pm 1$ it coincides with the spherical reflector as expected. Also note that the symmetry or focal plane ($\xi = 0$) is a surface of constant phase.

The field distribution throughout the resonator is given by the modulus of (20). The complete field distribution within the confocal resonator is shown schematically in Fig. 5 for the low-loss TEM_{00q} mode.

The field distribution over the focal plane is less spread out than over the spherical reflectors. The field spot size over the spherical reflectors was defined by (19). In any arbitrary plane z_0 the exponential term in the field distribution falls to e^{-1} at a radius

$$w_s = \sqrt{\frac{b\lambda(1 + \xi^2)}{2\pi}}. \tag{24}$$

The smallest achievable spot size is in the focal plane at $\xi = 0$.

To obtain the radiation pattern angular beam width of the TEM_{00q} mode spherical wave, one takes the ratio of the spot diameter from (20) or (24), as $\xi \to \infty$, to the distance from the center of the resonator. The beam width between the *half-power points* is given by

$$\theta = 2 \sqrt{\frac{\ln 2}{\pi}} \sqrt{\frac{\lambda}{b}} = 0.939 \sqrt{\frac{\lambda}{b}} \text{ radians}. \tag{25}$$

V. RESONATOR WITH NONCONFOCAL SPACING

Since the surfaces of constant phase of the confocal resonator are spherical, it is apparent that (20) also represents approximately the field distribution between two spherical reflectors of arbitrary spacing. That is, any two surfaces of constant phase may be replaced by reflectors. The frequencies at which such a resonator will be resonant will of course be determined by satisfaction of the phase condition.

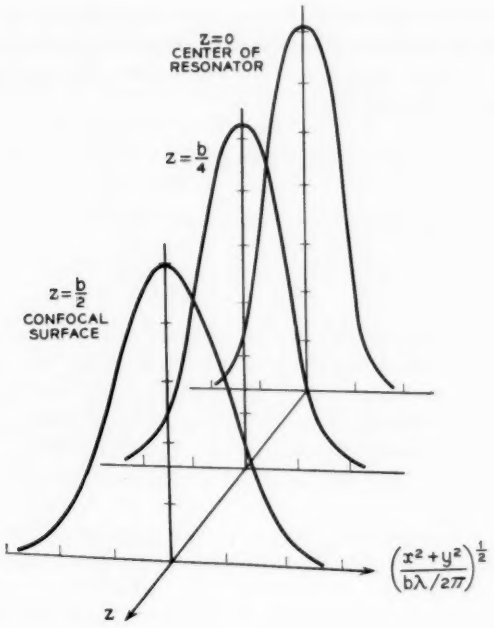


Fig. 5 — Field strength distribution within the confocal resonator for the TEM_{00q} mode.

Consider two identical spherical reflectors of radius of curvature b' spaced a distance d. The only restriction is that $b' \geqq d/2$. The confocal geometry of spacing b of which this resonator is a part is [set $\xi = d/b$ in (23)]:

$$b^2 = 2db' - d^2, \qquad b' \geqq \frac{d}{2}. \tag{26}$$

The spot size at the reflectors in the nonconfocal resonator may be immediately obtained from (24) with $\xi = \pm d/b$. It is

$$w_s' = \left(\frac{d\lambda}{\pi}\right)^{\frac{1}{2}} \left[2\frac{d}{b'} - \left(\frac{d}{b'}\right)^2\right]^{-\frac{1}{4}} \tag{27}$$

Note that the factor $[2(d/b') - (d/b')^2]$ achieves a maximum of unity, as a function of b', when $b' = d$. Thus, for a given spacing between reflectors, the spot size is a minimum for the confocal resonator.

One may estimate the loss of a nonconfocal resonator of square cross

section of dimension $2a'$ on the assumption that this loss is equal to that of its equivalent confocal resonator with reflector dimensions scaled up by the ratio of their spot sizes. The equivalent confocal resonator has spacing b, and its aperture is

$$2a = 2a' \frac{w_s}{w_s'} = 2a' \left(2 - \frac{d}{b'}\right)^{\frac{1}{4}}. \tag{28}$$

The important parameter in determining losses is

$$\frac{a^2}{b\lambda} = \frac{a'^2}{d\lambda}\left[2\frac{d}{b'} - \left(\frac{d}{b'}\right)^2\right]^{\frac{1}{2}}. \tag{29}$$

For given values of a' and d, the loss parameter is maximized, and thus losses are minimized, when $b' = d$. But this is just the confocal case. Thus the confocal geometry gives minimum spot size and minimum losses for a given spacing. If one defines the mode volume as the spot size at the reflector times the spacing, it is clear that the minimum mode volume also results from the confocal geometry. The mode volume, so defined, is

$$V = \pi w_s'^2 d = \lambda d^2 \left[2\frac{d}{b'} - \left(\frac{d}{b'}\right)^2\right]^{-\frac{1}{2}}. \tag{30}$$

It is important to note that the results of this section are valid only when the diffraction losses derived from the "equivalent" confocal geometry are small, that is, when the reflector dimension a' is somewhat larger than the spot size. In an exact solution for the nonconfocal case one should again start from the integral (4), and clearly the field distribution and losses so derived will depart from that obtained from the equivalent confocal case if the confocal field is not substantially all intercepted by the nonconfocal reflectors. Conversely, so long as the spot size is small compared to the reflector dimension a', one expects the field distribution and losses to be very nearly correctly given by the equivalent confocal solution.

The phase shift between the two reflecting nonconfocal surfaces may be obtained from (20). The condition of resonance may then be shown to be

$$\frac{4d}{\lambda} = 2q + (1 + m + n)\left(1 - \frac{4}{\pi}\tan^{-1}\frac{b - d}{b + d}\right). \tag{31}$$

In the nonconfocal case $4d/\lambda$ is no longer necessarily an integer at resonance. It is more important, though, that the modes are no longer degenerate in $m + n$. The spectral range or mode separation for the nonconfocal resonator is given by

$$\Delta\left(\frac{1}{\lambda}\right) = \frac{1}{4d}\left[2\Delta q + \left(1 - \frac{4}{\pi}\tan^{-1}\frac{b-d}{b+d}\right)\Delta(m+n)\right]. \quad (32)$$

Note that in the confocal case the set of modes $mnq = 00q, 01q$ are maximally split in frequency, whereas if the parameter in parenthesis equals $\frac{1}{2}$ (when $d/b = 0.414$) then the $mnq = 00q, 01q, 11q, 12q$ modes are maximally split in frequency.

When $b \approx d$, (31) becomes

$$\frac{4d}{\lambda} \approx 2q + (1+m+n)\left[1 - \frac{2}{\pi}\left(1 - \frac{d}{b}\right)\right], \quad (33)$$

where m and n are small integers and q a large integer. In the confocal case ($b = d$) note that equations (31) and (33) reduce to (14).

The theory of this section does not extend to the limit of plane-parallel reflectors, i.e., infinite radii of curvature. Let the spacing d remain fixed while b', and consequently [by (26)] the confocal radius b, approaches infinity. The spot size, as seen from (27), keeps increasing with b', and, as has been noted above, this results eventually in the breakdown of the whole idea of an equivalent confocal resonator. The relations for the nonconfocal resonator are valid as long as the reflector aperture radius a' is somewhat larger than the field spot size radius given by (27). That is, one must require

$$\frac{a'^2}{d\lambda} > \frac{1}{\pi}\left[2\frac{d}{b'} - \left(\frac{d}{b'}\right)^2\right]^{-\frac{1}{2}}. \quad (34)$$

VI. RESONANT MODES OF THE PLANE-PARALLEL RESONATOR

For comparison purposes, consider the resonances of a rectangular conducting box in the manner of Schawlow and Townes.[1] Let the dimensions be $2a \times 2a \times b$:

$$\left(\frac{2}{\lambda}\right)^2 = \left(\frac{q}{b}\right)^2 + \left(\frac{r}{2a}\right)^2 + \left(\frac{s}{2a}\right)^2, \quad (35)$$

where q, r, and s are integers. Modes where $q \gg r,s$ can be thought of physically as waves bouncing predominantly back and forth between the reflecting end plates of the rectangular box. The spectral range or mode separation is given by

$$\Delta\left(\frac{1}{\lambda}\right) = \frac{1}{2b}\left[\Delta q + \frac{1}{16}\left(\frac{b\lambda}{a^2}\right)(2r\Delta r + \Delta r^2 + 2s\Delta s + \Delta s^2)\right], \quad (36)$$

where $q \approx 2b/\lambda$ for $r, s = 1,2,3,\cdots$.

Removing the conducting side walls causes large diffraction losses for

the $r = 0$ or $s = 0$ modes since they have a strong field at the edge of the reflectors. Large r or s modes represent waves traveling at a considerable angle to the normal between the reflectors and thus these modes have such large diffraction losses that they are eliminated as resolvable resonant modes. Modes with $r, s = 1,2,\cdots$, have small diffraction losses, and are approximations to the actual modes which can exist in the resonator without the conducting side walls. Fox and Li's[6] work shows that for $a^2/b\lambda$ greater than unity the mode separations of a plane-parallel Fabry-Perot are given approximately by (36), the approximation improving rapidly with increasing $a^2/b\lambda$.

The mode separation corresponding to Δr or $\Delta s = 1$ has, to the writers' knowledge, never been resolved at optical frequencies due to the large values of $a^2/b\lambda$ and low values of reflectance used. Calculations show, though, that for reflectance coefficients of about 0.99 and $a^2/b\lambda \approx 4$, such that diffraction losses are comparable with reflector losses, the resonances should be resolvable.

The mode separation due to $\Delta q = 1$ is easily resolvable and is given by $\Delta(1/\lambda) = 1/2b$. This is the spectral range as normally stated for the plane parallel Fabry-Perot interferometer. It corresponds to changing the number of half wavelengths between the reflecting surfaces by one.

The confocal resonator is resonant for integer values of $4b/\lambda$. The mode separation due to $\Delta q = 1$ is $\Delta(1/\lambda) = 1/2b$. The modes are degenerate in frequency in that for a given integer $4b/\lambda$ all TEM_{mnq} modes are resonant such that $m + n$ remains even or odd according to whether $4b/\lambda$ is odd or even. The modes of the plane-parallel Fabry-Perot are not degenerate, except for rsq and srq. A possible advantage of this degeneracy of the confocal modes will be discussed in the next section.

VII. CONFOCAL RESONATOR APPLIED TO OPTICAL MASERS

A type of solid state optical maser has recently been demonstrated by Maiman[14] and by Collins et al.[15] It consists of a fluorescent crystal material (ruby) a few centimeters in length and a few millimeters in diameter. The crystal material should be optically homogeneous. The ends of the crystal are optically flat and parallel. The ends are silver-coated for high reflectance. One of the reflecting surfaces must be slightly transparent, as the output of the optical maser is obtained through the reflecting surfaces. Thus far, silver has been used to provide the reflection, but for ultimate performance multiple-layer dielectrics[5] should be used to obtain low transmission loss as well as high reflectance. The pump power enters the fluorescent crystal from the side.

It is seen in (2) that, if diffraction losses are small compared to re-

flection losses, then the resonator Q is proportional to the spacing between the reflecting surfaces. Consider a confocal resonator and a plane-parallel resonator each of spacing b and of equal Q. The energy distribution in the former is more concentrated on the axis and thus the confocal resonator has a smaller effective mode volume. The volume of maser material required will thus be less for the confocal than for the plane parallel resonator. For maser oscillation the required excess density of excited states depends only on the cavity Q and in no other way upon the resonator shape.[1] The pump power is proportional to the volume of maser material times the density of excited states divided by the natural lifetime of the excited state. Thus, assuming equal Q, the confocal resonator with its smaller volume of material requires less pump power than the plane parallel resonator by the ratio of their cross-sectional areas. Snitzer[16] recently pointed out this relation between mode volume and pump power with regard to the use of optical fibers in maser applications.

The minimum volume of maser material is limited by diffraction losses. If diffraction losses are to be considerably less than 1 per cent for the lowest-order mode so as to be small compared to achievable reflection losses, then $a^2/b\lambda \geqq 1$. The minimum volume of maser material is then

$$V_m = \pi a^2 b \approx \pi b^2 \lambda. \tag{37}$$

If $b = 4$ cm and $\lambda = 10^{-4}$ cm, then the rod of maser material should be approximately 0.4 mm in diameter. A rod of larger diameter would waste pump power in that the field of the confocal resonator would be very weak outside this minimum diameter of material.

The analysis of the confocal resonator assumes a uniform dielectric material between the spherical reflectors. For reasons of minimizing the pump power, it is necessary to use a small diameter of maser material. Therefore, to prevent internal reflection of energy from the sides of the maser material, it may be advisable to grind rough the sides of the rod of maser material or to immerse it in a surrounding medium of equal dielectric constant. If this is not done, the energy assumed lost due to diffraction effects would not escape and the electric field pattern will not be as computed herein. A more important effect of internal reflection from the side walls would be to increase the Q of the transverse modes which would increase the spontaneous and stimulated emission power to these undesired modes, and thus increase the over-all pump power required.

The natural linewidth of the material used in an optical maser will, for reflector spacing d of a few centimeters, be large compared to the

mode separation determined by integer changes in r,s for a plane-parallel resonator. Hopefully, the natural linewidth of the maser material will be less than the mode separation corresponding to integer changes in q. Thus, there is the possibility that a plane-parallel resonator optical maser may frequency wander between low-order r,s modes.

If the diffraction losses are comparable to or exceed the reflection losses for the lowest mode then, as can be seen from Fig. 3, the ratio of the Q's of the lowest two modes of the confocal resonator exceeds considerably the ratio of the Q's of the lowest two modes of the plane-parallel resonator. By the lowest order mode is meant $m = n = 0$, and $r = s = 1$, respectively, for the confocal and plane-parallel resonator. Therefore, maser oscillation is more likely to take place in *only* the lowest-order mode of the confocal than of the plane-parallel resonator. This greater loss discrimination between modes may be one of the significant advantages of the confocal resonator.

In the confocal resonator optical maser, if the maser oscillation wanders between modes the output beam pattern will change, just as in the plane-parallel resonator, but the frequency will remain fixed due to the mode degeneracy. Thus, the observed linewidth of the maser output may be narrower for the confocal resonator.

The required accuracy on the confocal condition to achieve degeneracy may be estimated from (33) and (36). It can be shown that if

$$| d/b - 1 | \approx 0.03 \qquad \text{and} \qquad (a^2/b\lambda) \approx 10,$$

the mode splitting of the near-confocal resonator equals the mode separation of the plane parallel resonator. To achieve a significantly smaller mode separation in the near-confocal resonator than the plane parallel resonator would require proportionately greater accuracy in the radius of curvature and spacing of the curved surfaces.

The plane Fabry-Perot requires accurately parallel reflecting surfaces. The confocal resonator requires only that the axis of the confocal resonator approximately coincide with the axis of the rod of maser material. The axis of the confocal resonator is the line passing through the two centers of curvature. The resonator axis must intersect the two reflecting surfaces near their center. Define the effective aperture radius as the distance from the point of intersection of the axis of the confocal resonator with the reflector surface to the nearest edge of the aperture. The diffraction losses will be approximately determined by this distance.

If the minimum diameter of maser material is used, then the axis of the confocal resonator must coincide with the material axis. Increasing

the diameter of the maser material wastes pump power but relaxes the tolerance on the resonator axis.

It is well to note that a single spherical reflecting surface and a plane reflecting surface spaced by approximately half the radius of curvature will have similar properties to the confocal resonator and may be advantageous if it is desired to bring the output through a plane surface.

VIII. CONCLUSIONS

A confocal multimode resonator formed by two spherical reflectors spaced by their common radii of curvature has been considered. The mode patterns and diffraction losses have been obtained. The confocal spacing of the reflectors is found to be optimum in the sense of minimum diffraction losses and minimum mode volume.

The diffraction losses are found to be orders of magnitude smaller than those of the plane-parallel Fabry-Perot, as obtained by Fox and Li.[6] It is more important, though, that a greater diffraction loss discrimination between modes occurs, and thus oscillation in other than the lowest-order mode is less likely for the confocal resonator, assuming that diffraction losses are comparable to reflection losses.

The modes of the confocal resonator are degenerate, in that one-half of all the possible field pattern variations over the aperture are resonant at any one time. This degeneracy is split if the resonator is nonconfocal. The splitting is comparable with that of the plane-parallel resonator (with $a^2/b\lambda \approx 10$) if the spacing of the reflectors is about 3 per cent different from the common radius. The mode volume and diffraction losses are insensitive to the confocal condition.

The required volume of maser material is smaller for the confocal resonator than for the plane-parallel resonator, and thus the required pump power is less. The confocal resonator is relatively easy to adjust in that no strict parallelism is required between the reflectors. The only requirement is that the axis of the confocal resonator intersect each reflector sufficiently far from its edge so that the diffraction losses are not excessive.

The example of a confocal resonator mentioned here was taken at infrared-optical wavelengths; however, such resonators may be useful down to the millimeter wave range by virtue of their low loss. In this connection, recent work of Culshaw[17] on the plane-parallel Fabry-Perot at millimeter wavelengths is of importance.

The writers have been informed that Goubau and Schwering[18] have recently investigated diffraction losses of parabolic reflectors and that their results agree with the work presented here.

IX. ACKNOWLEDGMENT

Fruitful discussions with A. G. Fox, W. D. Lewis, T. Li, D. Marcuse, S. P. Morgan and G. W. Series are sincerely appreciated. Mrs. F. J. MacWilliams performed the computations.

REFERENCES

1. Schawlow, A. L. and Townes, C. H., Phys. Rev., **112,** 1958, p. 1940.
2. Prokhorov, A. M., J.E.T.P., **34,** 1958, p. 1658.
3. Dicke, R. H., U.S. Patent 2,851,652, September 9, 1958.
4. Meissner, K. W., J. Opt. Soc. Am., **31,** 1941, p. 405; **32,** 1942, p. 185.
5. Heavens, O. S., *Optical Properties of Thin Films*, Butterworths, London, 1955.
6. Fox, A. G. and Li, T., this issue, p. 453; Proc. I.R.E., **48,** 1960, p. 1904.
7. Lewis, W. D., private communication.
8. Connes, P., Revue d'Optique, **35,** 1956, p. 37; J. Phys. Radium, **19,** 1958, p. 262.
9. Jenkins, F. A. and White, H. E., *Fundamentals of Optics*, 3rd ed., McGraw-Hill, New York, 1957.
10. Silver, S., *Microwave Antenna Theory and Design*, M.I.T. Radiation Laboratory Series, Vol. 12, McGraw-Hill, New York, 1949.
11. Slepian, D. and Pollak, H. O., B.S.T.J., **40,** 1961, p. 43.
12. Flammer, C., *Spheroidal Wave Functions*, Stanford Univ. Press, Palo Alto, Calif., 1957.
13. Slepian, D. and Pollak, H. O., private communication.
14. Maiman, T. H., Nature, **187,** 1960, p. 493.
15. Collins, R. J., Nelson, D. F., Schawlow, A. L., Bond, W., Garrett, C. G. B. and Kaiser, W., Phys. Rev. Letters, **5,** 1960, p. 303.
16. Snitzer, E., J. Appl. Phys., **32,** 1961, p. 36.
17. Culshaw, W., I.R.E. Trans., **MTT-7**, 1959, p. 221; **MTT-8**, 1960, p. 182.
18. Goubau, G. and Schwering, F., U.R.S.I.-I.R.E. Spring Meeting, Washington, May 1960.

Relation Between Surface Concentration and Average Conductivity in Diffused Layers in Germanium

By D. B. CUTTRISS

(Manuscript received July 18, 1960)

In this paper an expression is derived for calculating the average conductivity of a diffused layer in semiconductor material as a function of the surface concentration of the diffused impurity and the background impurity concentration. Curves are presented depicting the relationship among these parameters for the case of germanium. Included are curves for both diffused impurity types for the complementary error function, gaussian, exponential and linear impurity distributions.

I. INTRODUCTION

In the design of semiconductor devices in which junctions are produced by solid state diffusion of impurities, it is of great value to know the relationships which exist between the surface concentration of the diffused impurity, C_0, the background impurity concentration, C_B, and the average conductivity of the diffused layer, $\bar{\sigma}$. These relationships can be calculated from a knowledge of the resistivity as a function of impurity concentration for material uniformly doped with a single impurity. Such calculations are presented in this paper.

II. DERIVATION OF THE AVERAGE CONDUCTIVITY EXPRESSION

For convenience, assume initially that impurity atoms are 100 per cent ionized. Therefore, the conductivity at a point in a diffused layer in semiconductor material can be given by

$$\sigma = q\mu\,(C - C_B), \tag{1}$$

where

q = electronic charge,
C = diffused impurity concentration,

C_B = background impurity concentration, and
μ = majority carrier mobility.

This expression is valid for $(C - C_B) \gg n_i$ so that minority carrier concentration is negligible. Also, for values of $C_B > 10^{14}$ mobility is primarily a function of the total number of ionized impurities present. If it is assumed that both ionized impurity types scatter a majority carrier identically, then the mobility in (1) may be considered to be a function of $(C + C_B)$.

The conductivity of material doped with a single impurity can be expressed, again assuming 100 per cent ionization of impurities, as

$$\sigma^* = q\mu N, \tag{2}$$

where N is the impurity concentration. Rewriting (1) as

$$\sigma = q\mu(C + C_B)\left(\frac{C - C_B}{C + C_B}\right) \tag{3}$$

and substituting (2) with $N = (C + C_B)$, results in

$$\sigma = \sigma^*\left(\frac{C - C_B}{C + C_B}\right). \tag{4}$$

A log-log plot of the resistivity of single impurity doped material as a function of the impurity concentration can be approximated by a set of intersecting straight lines each having an equation of the form

$$\rho = \frac{1}{\sigma^*} = \frac{1}{B}N^{-\alpha} \tag{5}$$

each of which is valid over a certain range of N. Substituting (5) in (4), again with $N = (C + C_B)$, one has

$$\sigma = B(C + C_B)^{\alpha}\left(\frac{C - C_B}{C + C_B}\right) \tag{6a}$$

$$= B(C + C_B)^{\alpha-1}(C - C_B). \tag{6b}$$

The average conductivity of a diffused layer may be obtained by integrating (6b) over values of x from the surface $(x = 0)$ to the junction $(x = x_j)$ and dividing by the junction depth, x_j. Thus,

$$\bar{\sigma} = \frac{1}{x_j}\int_0^{x_j} B[C(C_0, x) + C_B]^{\alpha-1}[C(C_0, x) - C_B]\,dx. \tag{7}$$

The values assigned to B and α at any point on the interval are determined by the value of $[C(C_0, x) + C_B]$ at that point. Equation (7)

generally requires numerical integration, using experimental values for B and α.

III. APPROXIMATIONS TO RESISTIVITY CURVE

Fig. 1 shows the variation of resistivity, ρ, of single-impurity doped germanium as a function of the impurity concentration, N. Points in the range of $10^{14} \leqq N \leqq 2 \times 10^{16}$ for n-type material and in the range of $10^{14} \leqq N \leqq 6 \times 10^{16}$ for p-type material were taken from Prince.[1] Points in the range of $2 \times 10^{16} \leqq N \leqq 10^{20}$ for n-type material and $6 \times 10^{16} \leqq N \leqq 10^{20}$ for p-type material were taken from Hall effect measurements of Tyler and Soltys.[2] Hall effect measurements give the resistivity as a function of carrier concentration. However, direct measurements of resistivity as a function of impurity concentration by Trumbore and Tartaglia[3] for p-type material agree with the results of Tyler and Soltys, thus justifying the assumptions made, at least for the case of p-type material. Five straight-line approximations were made to

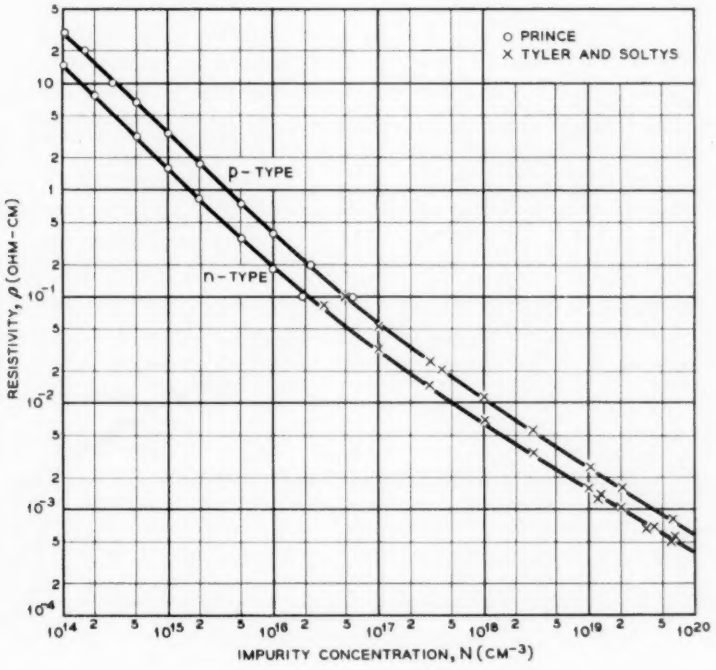


Fig. 1 — ρ vs. N for germanium at 300°K.

TABLE I — CONSTANTS FOR APPROXIMATIONS TO ρ VS. N CURVE

$$\rho = \frac{1}{B} N^{-\alpha}$$

Range	B	α
	n-type	
$10^{14} \leqslant N \leqslant 10^{15}$	2.74×10^{-15}	0.957
$10^{15} \leqslant N \leqslant 10^{16}$	1.22×10^{-14}	0.914
$10^{16} \leqslant N \leqslant 10^{17}$	5.05×10^{-13}	0.813
$10^{17} \leqslant N \leqslant 10^{18}$	1.70×10^{-10}	0.664
$10^{18} \leqslant N \leqslant 10^{20}$	1.74×10^{-9}	0.608
	p-type	
$10^{14} \leqslant N \leqslant 10^{16}$	1.61×10^{-15}	0.950
$10^{16} \leqslant N \leqslant 10^{17}$	6.28×10^{-14}	0.851
$10^{17} \leqslant N \leqslant 10^{18}$	1.74×10^{-11}	0.707
$10^{18} \leqslant N \leqslant 10^{19}$	1.61×10^{-10}	0.653
$10^{19} \leqslant N \leqslant 10^{20}$	1.11×10^{-9}	0.609

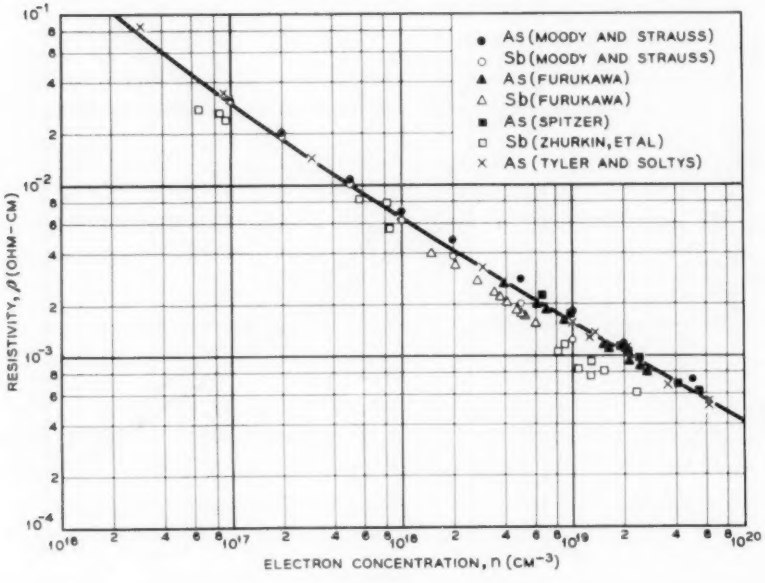


Fig. 2 — ρ vs. n for germanium at 300°K, from Hall effect measurements.

each curve giving equations of the form of (5). Values of B and α and the range of validity of each set of values are shown in Table I.

Fig. 2 shows data of resistivity as a function of electron concentration for n-type germanium as reported by Tyler and Soltys,[2] Moody and Strauss,[4] Furukawa,[5] Zhurkin et al.[6] and Spitzer.[7] Also shown in Fig. 2 is a portion of the n-type curve from Fig. 1. As can be seen, this curve represents a reasonable average of the arsenic data, for which case the present calculations are intended.

IV. RESULTS

Equation (7) was evaluated on the IBM 704 computer for various impurity distributions, and the results were checked by hand calculation of several points. Seven values of background concentration were used, and four points per decade of surface concentration were evaluated. The results are shown graphically in Figs. 3 through 10 on pages 514 through 521 for the various distributions as follows:

1. Complementary error function, Figs. 3 and 7.
2. Gaussian, Figs. 4 and 8.
3. Exponential, Figs. 5 and 9.
4. Linear, Figs. 6 and 10.

Since the conductivity of perfectly compensated material is not zero, (7) will be in error by some small amount. However, for values of C_0 and C_B such that $(C_0 - C_B) \geqq 10\ n_i$ this error will be negligible. All values of C_0 and C_B used in these calculations fulfill this requirement.

V. ACKNOWLEDGMENTS

The author wishes to thank J. C. Irvin and W. H. Jackson for many helpful and enlightening discussions, and special thanks are due Mrs. G. N. Alfandre whose able programming of the IBM 704 made solution of (7) possible.

REFERENCES

1. Prince, M. B., Drift Mobilities in Semiconductors — I. Germanium, Phys. Rev., **92,** 1953, p. 681.
2. Tyler, W. W. and Soltys, T. J., General Electric Research Lab. Memo Report No. P-193.
3. Trumbore, F. A. and Tartaglia, A. A., Resistivities and Hole Mobilities in Very Heavily Doped Germanium, J. Appl. Phys., **29,** 1958, p. 1511.
4. Moody, P. L. and Strauss, A. J., Electrochemical Society Meeting, Chicago, May 2, 1960; also private communication.
5. Furukawa, Y., Tunneling Probability in Germanium p-n Junctions, J. Phys. Soc. Japan, **15,** 1960, p. 730.
6. Zhurkin, B. G., et al., Izvestia Akad. Nauk. SSSR, OTN, MiT, No. 5, 1959, p. 86.
7. Spitzer, W. G., private communication.

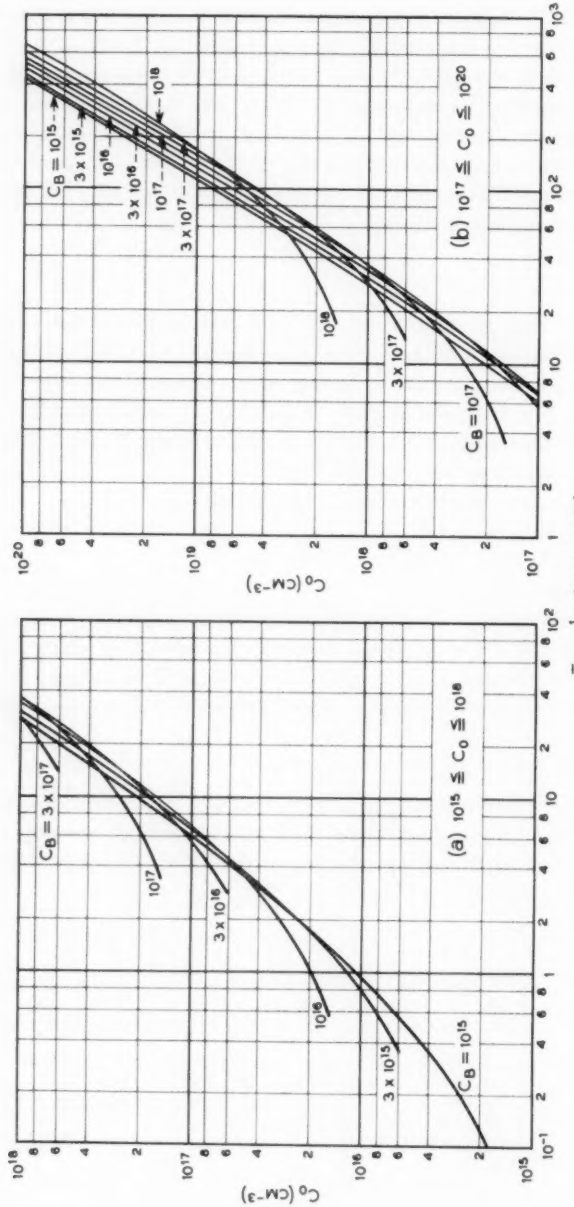


Fig. 3 — C_0 vs. $\bar{\sigma}$, p-type diffusion, ERFC distribution.

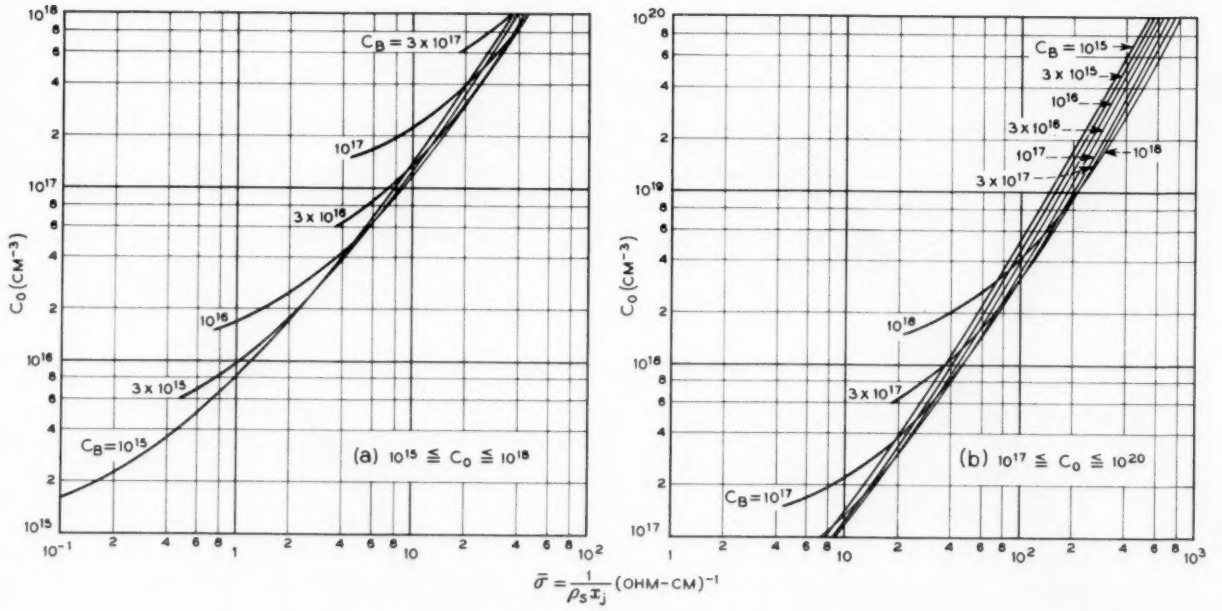


Fig. 4 — C_0 vs. $\bar{\sigma}$, p-type diffusion, gaussian distribution.

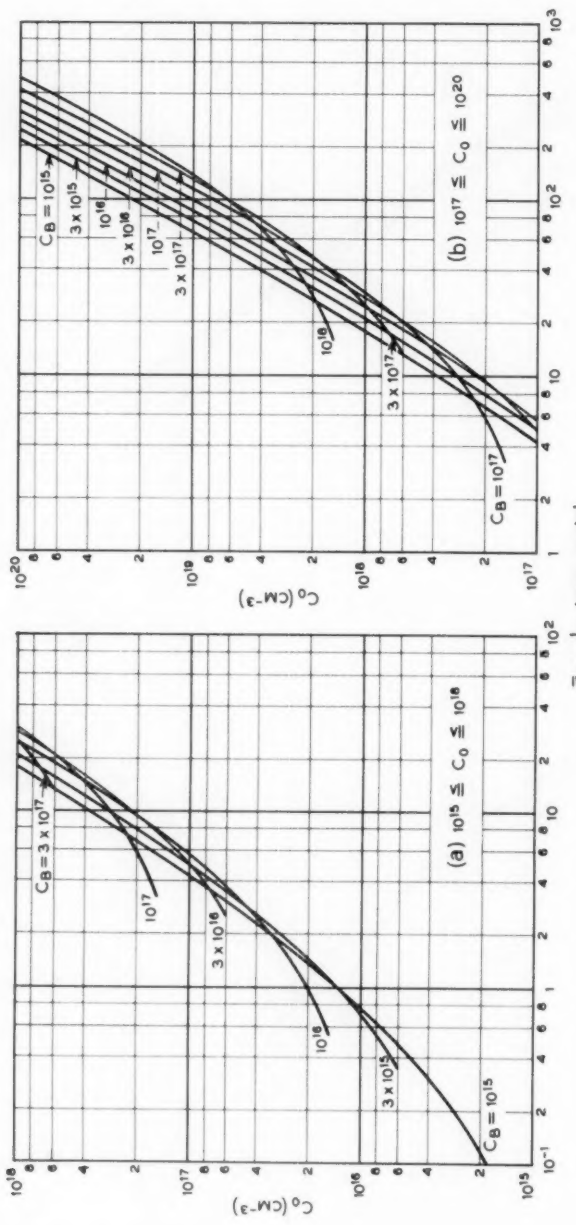


Fig. 5 — C_0 vs. $\bar{\sigma}$, p-type diffusion, exponential distribution.

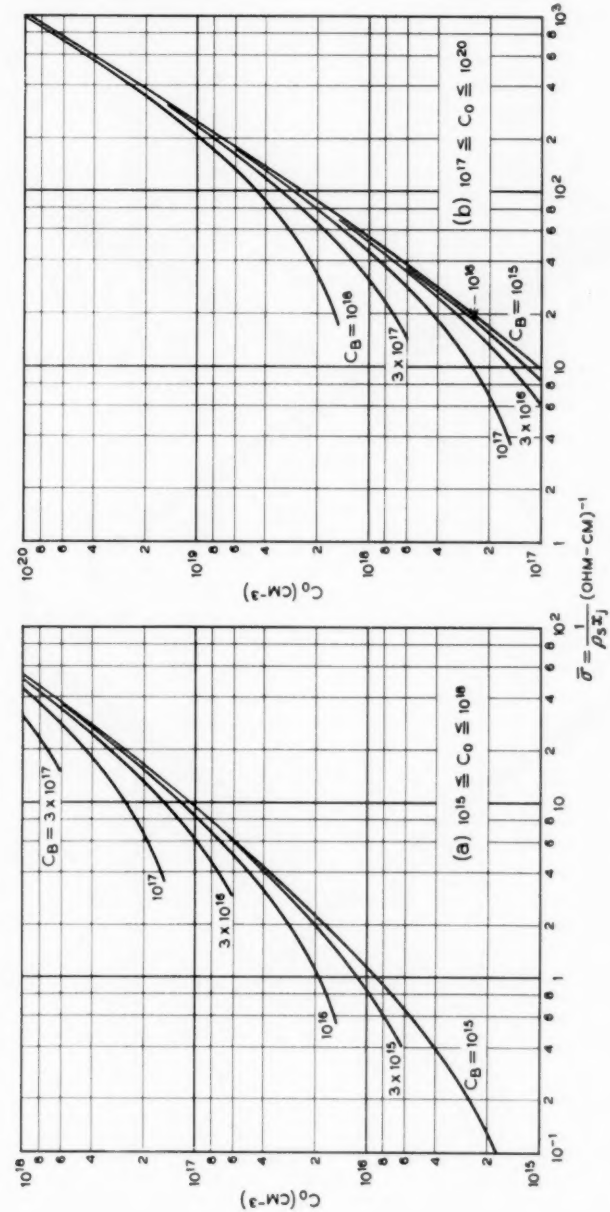


Fig. 6 — C_0 vs. $\bar{\sigma}$, p-type diffusion, linear distribution.

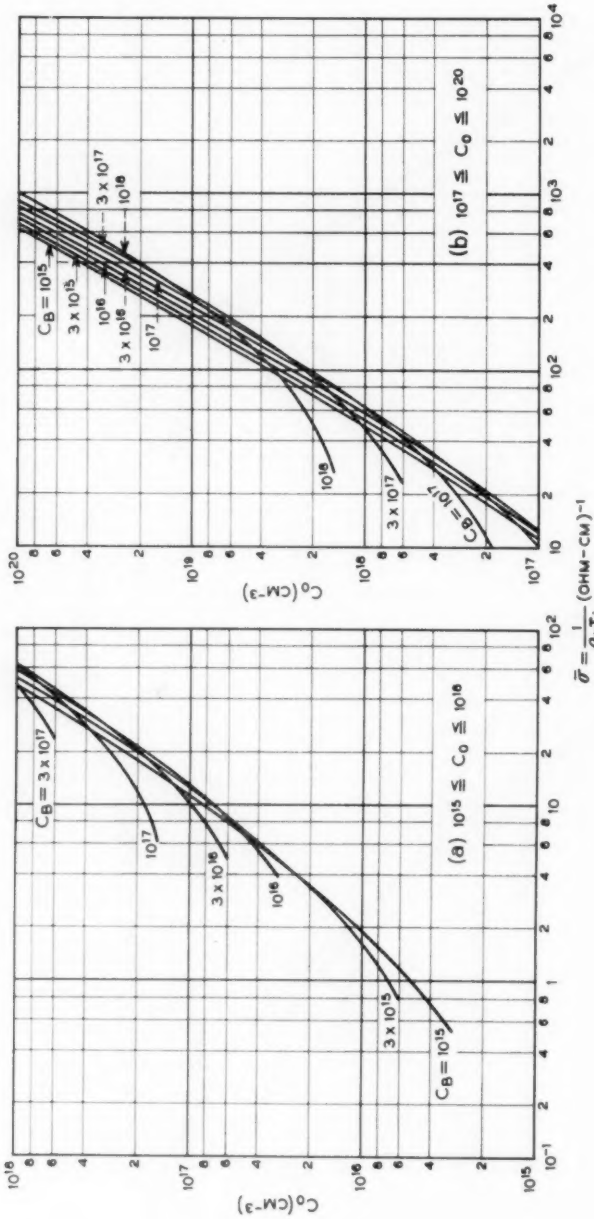


Fig. 7 — C_0 vs. $\bar{\sigma}$, n-type diffusion, ERFC distribution.

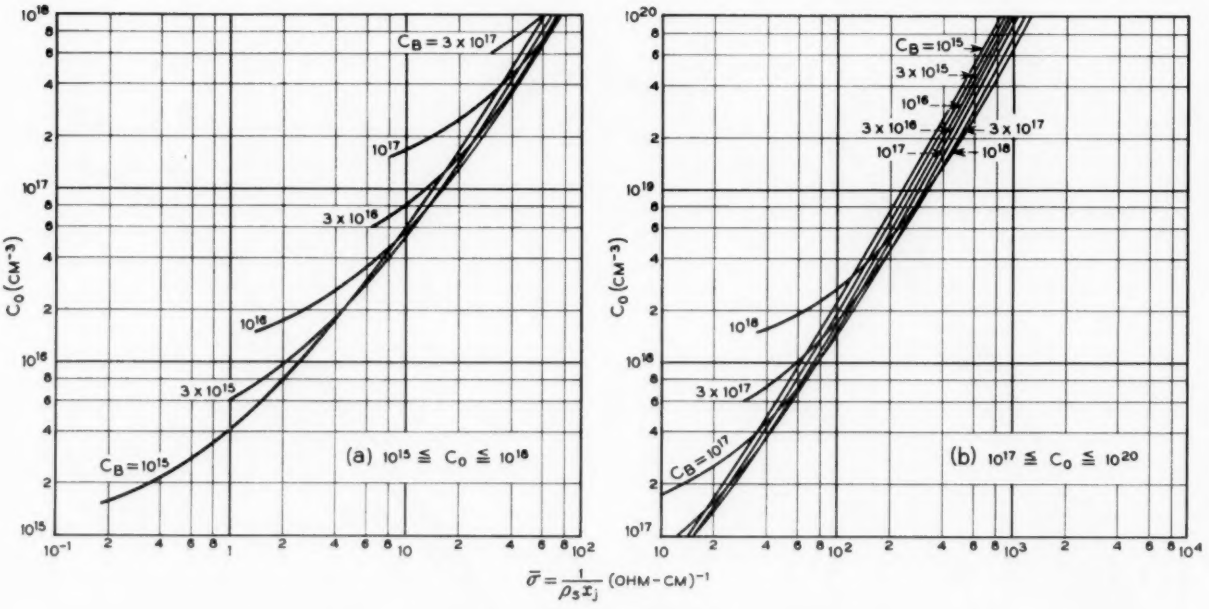


Fig. 8 — C_0 vs. $\bar{\sigma}$, n-type diffusion, gaussian distribution.

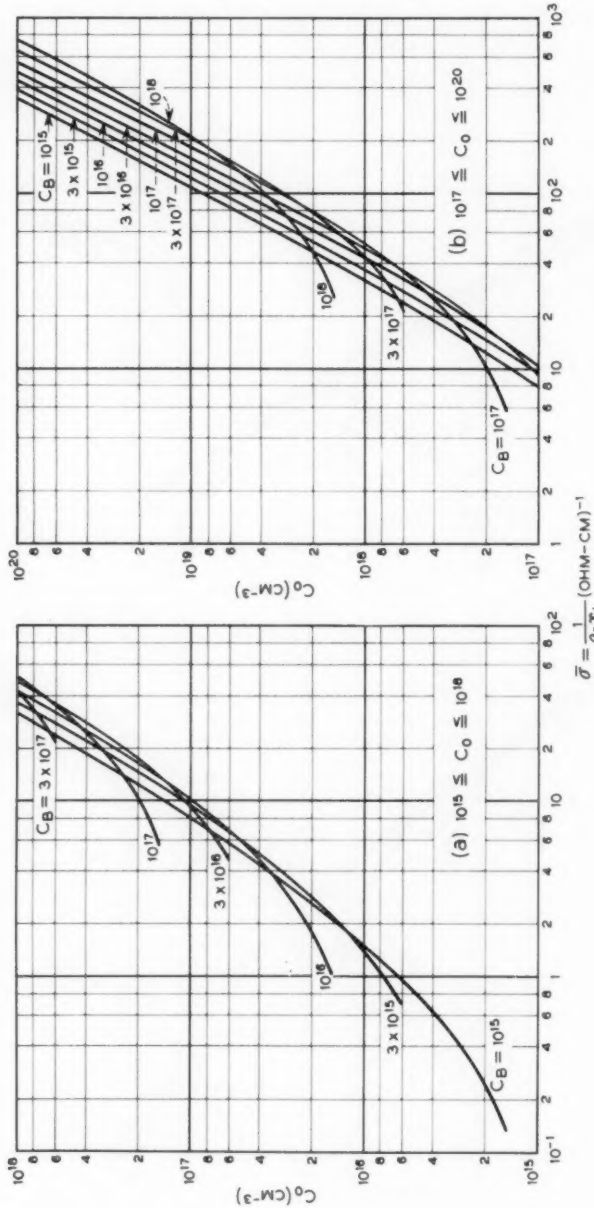


Fig. 9 — C_0 vs. $\bar{\sigma}$, n-type diffusion, exponential distribution.

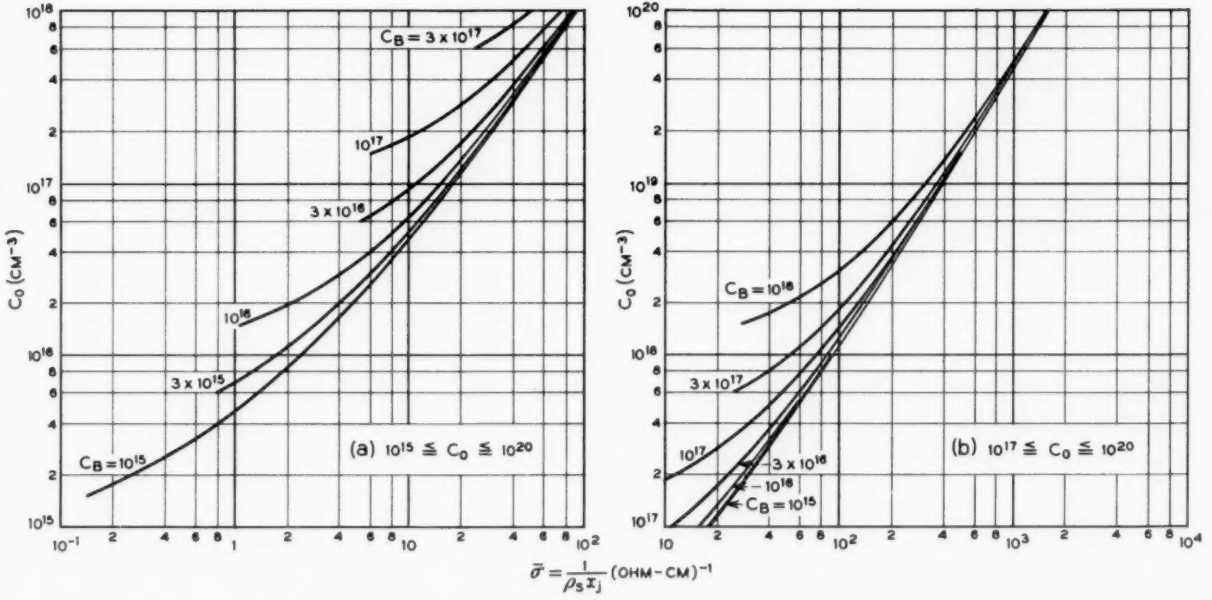


Fig. 10 — C_0 vs. $\bar{\sigma}$, n-type diffusion, linear distribution.

Magnetization and Pull Characteristics of Mating Magnetic Reeds

By R. L. PEEK, JR.

(Manuscript received June 8, 1960)

Expressions are presented for the magnetization and pull relations of the mating flat magnetic reeds used as contacting members in reed relays. The results of an experimental study expressed in dimensionless form give the force or pull between the reeds in terms of their dimensions, the gap between them, and the flux density. Since the attainable flux density is limited by saturation, the pull expression leads to conditions which must be satisfied by the reed and gap dimensions to provide desired levels of contact and retractile force.

The ampere-turn sensitivity of a relay using mating magnetic reeds depends on the flux required and on the reluctance of the magnetic path through the reeds. Expressions are given for the reluctance in the case of an air return path and in that where the air return is partially replaced by a shielding member.

Expressions are also given for the operate time of such reed relays. This depends on the concurrent flux development and motion of the reeds. The time of flux development varies inversely with the power input to the coil, but the motion time cannot be less than that required when the flux density is raised abruptly to its maximum value at the start of motion.

I. INTRODUCTION

A sealed reed relay comprises one or more sealed contacts assembled with a coil and shielding and supporting members. The distinctive feature is the construction of a sealed contact.[1,2] In its simplest form, as shown in Fig. 1, it consists of a glass envelope in which are sealed a pair of magnetic reeds, which serve as contacting members, actuated by the magnetic field induced in them by the coil current.

In another form,[2] there is only one moving reed, the other magnetic member being short and nearly rigid. This construction, using two of these short members to provide both back and front contacts, has been

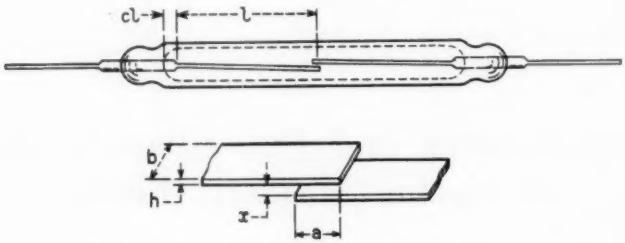


Fig. 1 — Dimensions of mating magnetic reeds.

used with mercury-wetted contact surfaces.[3,4] Both forms have been used in single- and multiple-contact relays, both neutral and with permanent magnet bias used to give polar or locking performance.[5] Recent work has included development of a miniature sealed contact.[6]

As an aid to the design of sealed contacts, and to the understanding of their characteristics, an analysis has been made of their performance relations, including the dependence of the pull, contact force, speed and sensitivity on the dimensions of the reeds. This analysis is presented here in the form applying to the simple case of equal mating reeds, but is, with some modification, applicable to other forms of magnetic reed contacts.

The major controlling factor in the performance of sealed contacts is the flux-carrying capacity of the reeds, as limited by the saturation density of the reed material. The treatment given here therefore starts from the relation between the reed flux and the attractive force at the gap, which must deflect the reeds and provide the contact force. This relation is determined essentially by the reed dimensions alone. The sensitivity, as measured by the ampere turns for operation and release, depends upon both the flux and the reluctance of the magnetic circuit, and hence upon the dimensions of the coil and shielding members, as well as upon those of the reeds. Relations are given for the estimation of reluctance and hence of sensitivity. Finally, expressions are given for the estimation of the speed of operation, which depends upon the times of field development and of reed motion, with the latter establishing a lower limit to the attainable time of operation.

II. PULL RELATION

Using the notation of Fig. 1 for the reed dimensions, the attractive force F between the reeds is given by Maxwell's law as

$$F = \frac{\varphi_a^2}{8\pi ab}, \tag{1}$$

where φ_G is the flux in the gap where the reeds overlap. The total flux Φ is the sum of φ_G and the fringing flux which passes from one reed to the other by air paths around the gap. As shown below in the discussion of reluctance, approximate estimates may be made of the fringing flux, but the relations involved are not adapted to a simple direct treatment of the relation between the pull F and the total flux Φ. A direct experimental study was therefore made to determine the approximate form of this relation.

In this study, pull and flux measurements were made of four different sets of reeds, having the dimensions listed in Table I. Each set of reeds was assembled with overlap values, a, of 25, 50 and 100 milli-inches, and the pull and flux measured for values of gap x in the range from 1 to 10 milli-inches over the range of coil energization from 0 to 300 ampere turns.

These measurements were all made with the reeds supported in a brass fixture which could be adjusted to make the overlap surfaces parallel and to set the overlap and the gap at desired values. The coil used had inside and outside diameters of 0.39 and 0.86 inch, and was 1.64 inches long. It was centered over the gap in making the measurements, and was provided with a central search coil located on its inner diameter. Measurements were also made of the flux in the coil alone, and this air-core flux was subtracted from the flux readings to give the reed flux Φ.

If the reeds are long compared with the overlap, as in practice and in these measurements, the pull is independent of the reed length. If the permeability of the reed material is high enough for the reeds to be essentially equipotential surfaces near the gap, the pull F is a function of Φ, x, a, b and h only. For dimensional consistency, therefore, the pull must be given by an equation of the form

$$\frac{8\pi abF}{\Phi^2} = f\left(\frac{x}{a}, \frac{a}{b}, \frac{h}{b}\right). \tag{2}$$

This is a convenient dimensionless form, since the left-hand term must, from (1), approach unity for $x = 0$. (In evaluating this term, consistent

TABLE I — DIMENSIONS OF REEDS TESTED

Thickness, h (milli-inches)	Width, b (milli-inches)	Length (inches)
10	100	1.75
20	100	1.75
30	100	1.75
21	60	1.75

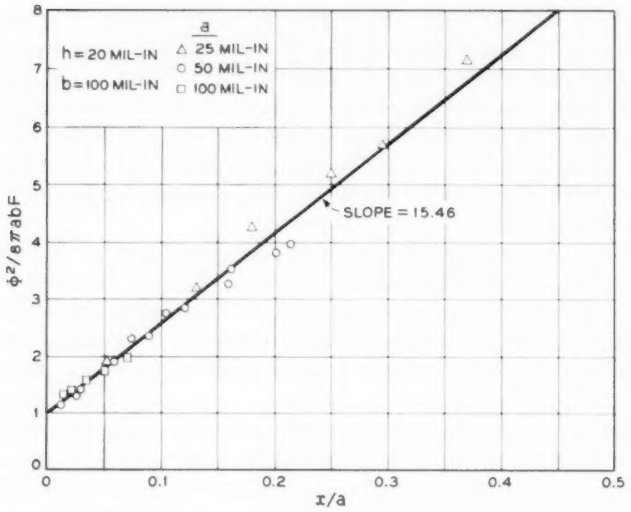


Fig. 2 — Results of pull measurements.

units must be used: in C.G.S. units, Φ is in maxwells, a and b in centimeters, and F in dynes.) From direct plots of the measured values of F and Φ there were read values of F for $\Phi = 10{,}000\ bh$ (i.e. for a flux density of 10,000 gauss). For each set of reeds and each overlap, the corresponding values of $\Phi^2/8\pi abF$ were plotted against x/a, as illustrated in Fig. 2, which shows the results for $h = 10$ milli-inches, $b = 100$ milli-inches and three values of a. As the points for all three values of a fall on the same curve, the right-hand side of (2) is independent of a/b, and reduces to a function of x/a and h/b only. The experimental relation of Fig. 2 is linear, and similar linear plots, independent of a/b, were obtained for the other three sets of reeds. Thus the functional relation (2) was found to be of the form

$$\frac{\Phi^2}{8\pi abF} = 1 + k\,\frac{x}{a}, \tag{3}$$

where k is a function of h/b. The observed values of k for the four sets of reeds were plotted against the corresponding values of h/b and a linear plot was obtained, conforming to the equation

$$k = 6.66 + 44.4\,\frac{h}{b}. \tag{4}$$

These results were obtained for values of Φ corresponding to a density

of 10,000 gauss. Similar plots were made for other values of density, and substantially the same relation was found to apply for densities ranging from 20 per cent to 80 per cent of the saturation density, which was about 15,000 gauss for the material used in these tests.

Within this range of reed flux density, therefore, all the observed values of pull conformed approximately to (3), with k given by (4). These expressions may therefore be used to estimate the pull of mating magnetic reeds at densities up to 80 per cent of saturation, when each reed is long compared with the overlap. The expressions are approximate, and minor deviations, particularly in the value of k, are produced by changes in coil length and in the return path and shielding configuration.

A further, and more important, deviation from these relations occurs in the closed gap condition, $x = 0$, where the pull is given by (3) as $\Phi^2/8\pi ab$. For release, this pull equals the retractile force of the reeds, shown in Fig. 3 as sX. Thus the flux at which release occurs should be given by $\sqrt{8\pi absX}$. Observed values of release flux show considerable variation, with the upper limit close to this computed value. This corresponds to the fact that surface irregularities or lack of parallelism of the mating surfaces concentrate the closed gap flux over a smaller area than ab, and increase the pull over that given by (3) for $x = 0$. The actual closed gap pull is therefore variable, and in general higher than the computed closed gap pull.

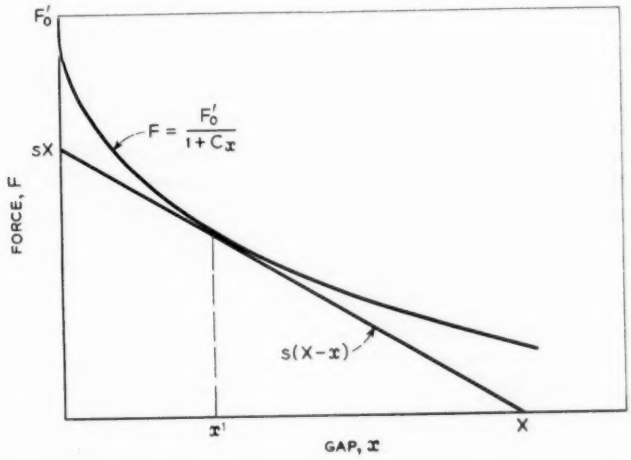


Fig. 3 — Pull and retractile force relations.

III. CAPABILITY

For operation, the pull curve of the reeds must exceed their stiffness load, shown in Fig. 3 as the force $s(X - x)$. Here X is the open gap separation, and s is the combined stiffness, or half the stiffness of one reed if the two are alike. (For unequal reeds, $1/s$ is the sum of the compliances of the two reeds.) Let F_0' be the closed gap pull, $\Phi^2/8\pi ab$, for the minimum value of Φ required for operation. Then from (3) the pull for this value of Φ is given by

$$F = \frac{F_0'}{1 + Cx}, \tag{5}$$

where $C = k/a$. At the point of tangency, $x = x'$, the pull and load curves must be equal and have equal slopes, so that:

$$\frac{F_0'}{1 + Cx'} = s(X - x'),$$

$$\frac{CF_0'}{(1 + Cx')^2} = s.$$

From these two conditions:

$$\frac{x'}{X} = \frac{CX - 1}{2CX}, \tag{6}$$

and, if $CX > 1$,

$$\frac{F_0'}{sX} = \frac{(1 + CX)^2}{4CX}. \tag{7}$$

If $CX \leqq 1$, the tangency condition no longer applies, the pull and load are equal at $x = 0$, and $F_0' = sX$.

After operation, the flux exceeds the just-operate value. Let F_0 be the value of the closed gap pull $\Phi^2/8\pi ab$ for the final operated flux. Then the contact force is equal to $F_0 - sX$, and the retractile force, tending to open the contact on release, is sX. The contact behavior varies with the contact force, so that one design requirement is that the contact force should exceed some specified minimum value F_C.

A second requirement is that the ratio of retractile force to the contact force should exceed some minimum value n (of the order of unity) to minimize the danger of sticking. The area of actual intimate contact, and hence the force required to break cold welds over this area, varies directly with the contact force. In formulating this requirement, the operated pull is taken as given by F_0. As stated above, the actual op-

erated pull is variable and in general exceeds F_0. Hence the actual contact force is usually in excess of its computed value, and the actual value of the retractile force ratio n less than its computed value.

A satisfactory design must meet the requirements for operation, contact force and retractile force through the range of dimensional variation occurring in manufacture. The present discussion is confined to the case where only variations in the gap X need be considered. Let the minimum and maximum gap values be X_1 and X_2, and let $c_1 = X_2/X_1$. The final flux must be below saturation, and for purposes of estimation may be assumed to correspond to a density B'', of the order of 90 per cent of the saturation density. For operation to occur without excessive reed reluctance, the operating density B' should be of the order of 80 per cent of saturation. Let $c_2 = F_0/F_0' = (B''/B')^2$. Then the operate condition (7) will be satisfied for all values of X if

$$\frac{F_0}{sX_2} = \frac{c_2(1 + CX_2)^2}{4CX_2}.$$

The contact force requirement is satisfied for all values of X if

$$F_0 = F_c + sX_2.$$

The retractile force requirement that $sX > n(F_0 - sX)$ is satisfied for all values of X if

$$\frac{F_0}{sX_1} = \frac{n + 1}{n}.$$

From the preceding three equations, the design requirements can be written as

$$\frac{4CX_2}{(1 + CX_2)^2} = \frac{c_1c_2n}{n + 1}, \tag{8}$$

$$F_0 = \frac{F_c}{1 - \dfrac{nc_1}{n + 1}}, \tag{9}$$

$$s = \frac{F_0 - F_c}{X_2}. \tag{10}$$

The reed dimensions must be chosen to satisfy these three equations for given values of X_1, c_1, F_c and n (c_2 being approximately fixed by the reed material used). The dimensions to be selected are a, b, h and whatever other dimension determines s: the free length l if the reeds are of uniform cross section. If the section ratio b/h is tentatively taken

as fixed by manufacturing considerations, the reed dimensions can be evaluated as follows:

The value of C is computed from (8), and since $a = k/C$ and k is given by (4), the required value of a is thereby determined.

Since the value of Φ for F_0 is bhB'',

$$F_0 = \frac{bh^2B''^2}{8\pi a}. \tag{11}$$

Then, with F_0 evaluated from (9), and a, B'' and b/h known, the thickness h (and hence the width b) can be determined from (11). The stiffness s is given by (10). For a reed of uniform section $b \times h$ the length l can be determined from the required value of s. As (11) determines the value of bh^2, and the stiffness varies as bh^3, the shortest reed meeting the requirements will be that for the largest value of b/h consistent with manufacturing considerations.

The relations outlined above allow for variations in the gap X, but not in the other dimensions. In practice, allowance must also be made for variations in the reed thickness h. This can be done by applying the treatment outlined above to the case where h has upper and lower limits, leading to expressions similar to (8), (9) and (10).

IV. ILLUSTRATIVE COMPUTATIONS

Using the relations given above, reed dimensions have been computed to give the three values of minimum contact force F_c shown as parameters for the curves of Fig. 4, over the range of gap dimensions shown. These illustrative cases were computed for the values of b/h, X_2/X_1, n and flux density shown in the legend of this figure. The values of flux density are those applying to iron-nickel alloys of about 50 per cent nickel. These have a Young's modulus value of about 25×10^6 psi, which was taken as applying in computing the length required to meet the stiffness requirement.

The computed dimensions shown in Fig. 4 are the reed thickness h, length l' and overlap a. The overlap varies inversely with the gap, as shown by (8) when b/h, n, c_1 and c_2 are fixed, as in these computed cases. The thickness and length both increase as the gap and the required contact force are increased. The gap is the major factor controlling the length and hence the over-all size of the sealed contact. The choice of gap is fixed in part by voltage breakdown requirements, and in part by manufacturing considerations which determine the variation in gap for which allowance must be made. The comparisons of Fig. 4 are somewhat idealistic in that a constant ratio c_1 ($= X_2/X_1$) is assumed. Actually,

the ratio c_1 is usually larger for small gaps than large ones, in which case the reduction in reed thickness and length resulting from reducing the gap is less than that shown in Fig. 4.

In the computations for Fig. 4 the retractile force ratio n is taken as unity. This is the computed minimum ratio of retractile force to contact

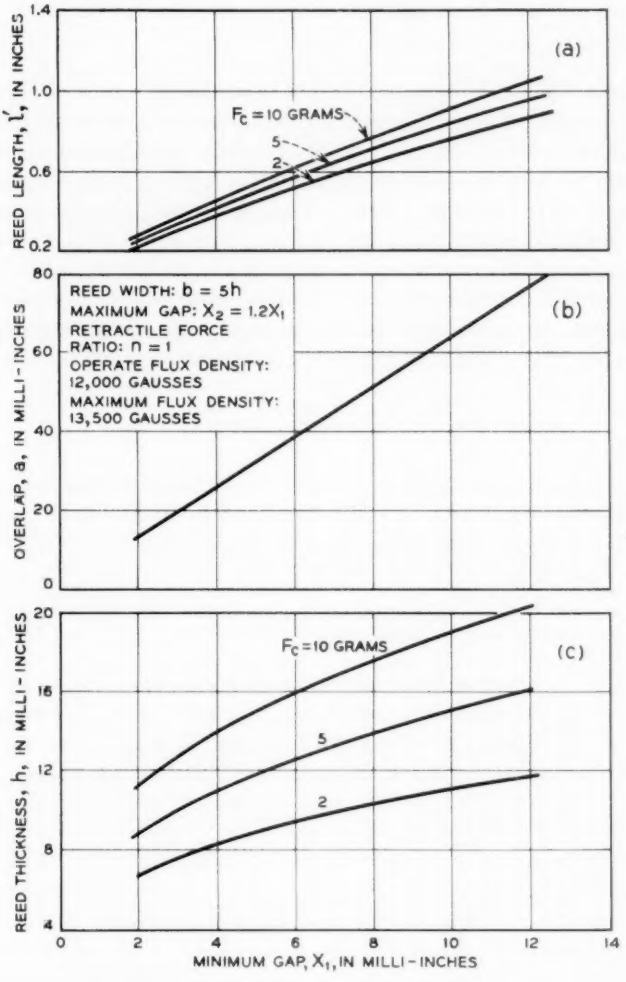


Fig. 4 — Illustrative computed reed dimensions.

force. As discussed previously, the actual contact force may be increased by flux concentration in the closed gap, and the actual ratio may be materially below the computed minimum, which serves therefore as a basis for comparing one design with another, rather than as an absolute measure of the liability to sticking. From (9), increasing n increases F_0, and thus increases the thickness h and the required length l'.

V. STIFFNESS ESTIMATION

The reed length is denoted l' in Fig. 4, as that giving the desired stiffness [$2s$ from (10)] for a uniform reed of length l' and stiffness given by $Ebh^3/4l'^3$. As shown in Fig. 1, convenient construction for the seal requires the use of a cylindrical section extending from the seal for a length cl, with the flat section extending beyond this for a length l. The cylindrical and rectangular sections are of equal area. The stiffness of such two section cantilevers are given in Fig. 6-30 of Peek and Wagar,[7] from which can be derived the following expression for the ratio l'/l of two cantilevers of equal stiffness, one having the dimensions shown in Fig. 1, and one of length l' and the same rectangular section as the other:

$$\left(\frac{l'}{l}\right)^3 = 1 + \left(\frac{\pi h}{3b}\right)[(1 + c)^3 - 1]. \tag{12}$$

Thus if cl and l are known, l' can be evaluated and used to determine the stiffness. Conversely, if l' is determined from the required stiffness, as in Fig. 4, the corresponding flat reed length l can be determined for a given value of cl.

VI. RELUCTANCE FOR AN AIR RETURN PATH

The flux Φ between the reeds required for operation can be estimated by means of the relations given in preceding sections. The corresponding value of coil magnetomotive force $\mathfrak{F}$, or $4\pi NI$, is given by $\mathfrak{R}\Phi$, where $\mathfrak{R}$ is the reluctance of the magnetic circuit. Estimating the sensitivity, or the value of $\mathfrak{F}$ for operation, therefore requires some method for estimating the value of the reluctance $\mathfrak{R}$.

Most reed relays have essentially an air return path, whose reluctance is only slightly reduced by the can or shield placed over the core. For an air return, the flux through the reeds is analogous to the current through a leaky transmission line, as discussed in Section 9-2 of Peek and Wagar.[7] Referring to Fig. 5, let φ be the flux in the reed at a distance x from the plane of symmetry, and let f be the potential difference

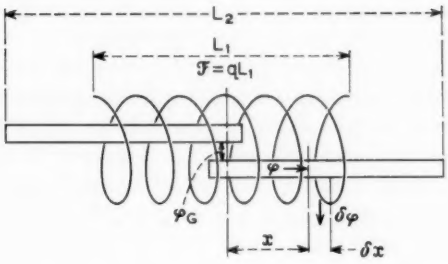


Fig. 5 — Notation for analysis of air return magnetic field.

between the point x and the plane of symmetry. The flux $d\varphi$ leaving the reed over the length dx is given by

$$-pf = \frac{d\varphi}{dx}, \tag{13}$$

where p is the permeance per unit length of reed of the air path from x to the plane of symmetry. The rate of change in f is given by

$$\frac{df}{dx} = q - r\varphi \qquad \text{for } x < \frac{L_1}{2},$$

$$\frac{df}{dx} = -r\varphi \qquad \text{for } \frac{L_1}{2} < x < \frac{L_2}{2},$$

where r is the reluctance of the reed per unit length and $q = \mathfrak{F}/L_1$, where L_1 is the coil length. Substituting (13) in these expressions gives the equations

$$\begin{aligned} \frac{d^2\varphi}{dx^2} - pr\varphi + pq = 0 \qquad & \text{for } x < \frac{L_1}{2}, \\ \frac{d^2\varphi}{dx^2} - pr\varphi = 0 \qquad & \text{for } \frac{L_1}{2} < x < \frac{L_2}{2}. \end{aligned} \tag{14}$$

Subject to limitations discussed below, p and r may be taken as constants, in which case the solution to (14) is given by

$$\begin{aligned} \varphi &= \frac{q}{r} + Ae^{\alpha x} + Be^{-\alpha x} \qquad \text{for } x < \frac{L_1}{2}, \\ \varphi &= Ce^{\alpha x} + De^{-\alpha x} \qquad \text{for } \frac{L_1}{2} < x < \frac{L_2}{2}, \end{aligned} \tag{15}$$

where

$$\alpha = \sqrt{pr}. \tag{16}$$

For continuity, the expressions for φ and $d\varphi/dx$ at $x = L_1/2$ given by the two forms of (15) must be equal, from which expressions are obtained for the coefficients C and D in terms of A and B. The coefficients A and B in turn are determined by the boundary conditions, which require that φ be zero at $x = L_2/2$ and that for a gap flux φ_G at $x = 0$, the potential $f = -\varphi_G\mathcal{R}_G/2$ at $x = 0$, where $\mathcal{R}_G$ is the gap reluctance. Then, from (13), the boundary conditions become

$$\varphi = \varphi_G\,, \quad \frac{d\varphi}{dx} = \frac{p\mathcal{R}_G\varphi_G}{2} \qquad \text{at } x = 0,$$

$$\varphi = 0 \qquad\qquad\qquad\qquad \text{at } x = \frac{L_2}{2}\,.$$

Substituting these conditions in (15), there are obtained

$$2A = \left(1 + \frac{p\mathcal{R}_G}{2\alpha}\right)\varphi_G - \frac{q}{r}\,, \tag{17}$$

$$2B = \left(1 - \frac{p\mathcal{R}_G}{2\alpha}\right)\varphi_G - \frac{q}{r}\,, \tag{18}$$

$$\varphi_G = \frac{q}{r}\,\frac{\cosh \alpha L_2/2 - \cosh \alpha(L_2 - L_1)/2}{\cosh \alpha L_2/2 + (p\mathcal{R}_G/(2\alpha)) \sinh \alpha L_2/2}\,. \tag{19}$$

As the boundary condition for $x = 0$ shows, the flux φ in the reed initially increases with x, and hence the point of maximum flux Φ is at some positive value of x where $d\varphi/dx = 0$. Applying this condition to the expression for φ given by (15), the maximum flux Φ is given by

$$\Phi = \frac{q}{r} - 2\sqrt{AB}.$$

This maximum flux Φ is the total flux between the reeds. The reluctance $\mathcal{R}$, or $\mathcal{F}/\Phi$, is given by qL_1/Φ. Substituting from the preceding equations, this expression for $\mathcal{R}$ may be written in the form

$$\mathcal{R}pL_2 = \frac{L_1}{L_2}\,\frac{(\alpha L_2)^2}{1 - \left[\left(1 - \dfrac{r}{q}\varphi_G\right)^2 - \left(\dfrac{Q}{2\alpha L_2}\,\dfrac{r}{q}\,\varphi_G\right)^2\right]^{\frac{1}{2}}}\,, \tag{20}$$

where $Q = \mathcal{R}_G pL_2$, and $r\varphi_G/q$ is given by (19). Thus the ratio $\mathcal{R}pL_2$ is a function of the three ratios L_1/L_2, αL_2 and Q, or $\mathcal{R}_G pL_2$. For the limit-

ing case where the reed reluctance per unit length r is so small that α approaches zero, the hyperbolic functions in (19) may be replaced by their series expansions and the higher power in α neglected. Then for $\alpha \to 0$, (20) reduces to the expression:

$$\Re pL_2 = \frac{\dfrac{2(4+Q)L_2}{2L_2 - L_1}}{1 + \dfrac{Q^2(2L_2 - L_1)L_1}{16(4+Q)L_2^2}}. \tag{21}$$

VII. DISCUSSION OF AIR RETURN RELUCTANCE

Fig. 6 shows $\Re pL_2$ for the case $L_1/L_2 = 0.5$ plotted against $\Re_G pL_2$ for various values of αL_2. The increase of $\Re$ with increasing αL_2 measures the effect of the reed reluctance. Aside from the reed reluctance, $\Re$ comprises the air return reluctance in series with the gap reluctance as shunted by the parallel air path across the gap. The curves show how this shunt path limits the increase in $\Re$ with increasing $\Re_G$.

The effect of increasing L_1/L_2 is to increase $\Re$, as can be seen from (21), except for large values of $\Re_G$ (and hence of Q). Thus the reluctance is reduced and the sensitivity increased by using a short coil concentrated over the gap.

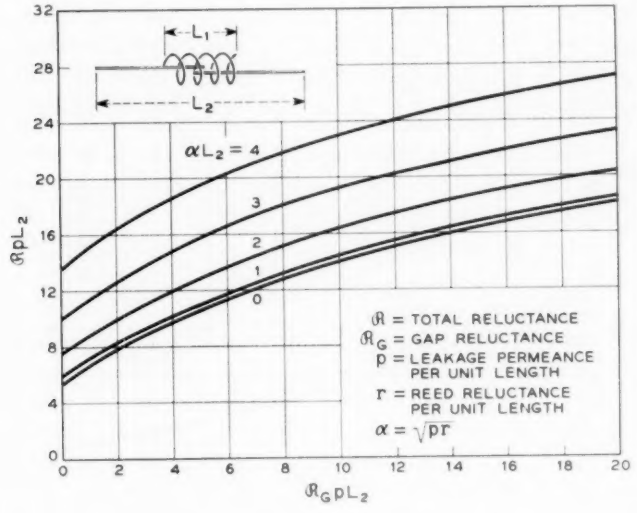


Fig. 6 — Reluctance relation for $L_1/L_2 = 0.5$.

In deriving the preceding equations, p and r have been taken as constants. The estimation of the air path permeance of bar magnets is discussed in a classic paper by Evershed,[8] who showed that satisfactory estimates can be obtained by using the constant value of p applying to a uniformly magnetized ellipsoid. This is the function of the ratio L/D shown in Fig. 7. In applying this to mating rectangular reeds, L is taken as the over-all length L_2, and D as $2(b + h)/\pi$.

The assumed constancy of r, the reluctance per unit length, is at best an approximation, since r is inversely proportional to the permeability of the material, which varies with the flux density, so that r varies along the length of the reeds. If, however, r is small, the error introduced by this approximation is minor.

The effect of the permeability on the reluctance and sensitivity, as related to the size and capability of mating reed contacts, is illustrated by the computed results of Table II. Here the dimensions h, b, a and l have been taken as those shown in Fig. 4 for minimum gap values of 4 and 10 milli-inches and contact force values of 2 and 10 grams. The operate flux values, as in Fig. 4, are for a density of 12,000 gauss. The over-all length L_2 has been taken as 3.75 times the computed reed length. The permeance p is taken from Fig. 7. The critical gap reluctance has been obtained by adding a 2 milli-inch air gap allowance for the closed gap to the critical gap x' computed from (6).

With $\mathcal{R}_6 pL_2$ thus computed for each of these two cases, values of αL_2 have been computed for two values of permeability μ, 1000 and 5000. With r given by $1/\mu bh$, corresponding values of αL_2 have been computed, and corresponding values of $\mathcal{R}pL_2$ read from Fig. 6, taking the coil

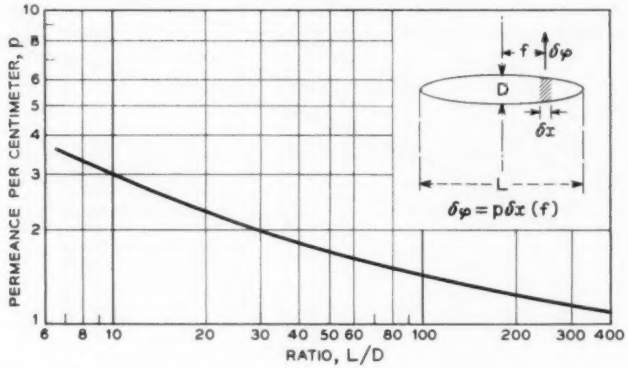


Fig. 7 — Permeance factor for ellipsoidal magnets.

TABLE II — ILLUSTRATIVE RELUCTANCE AND SENSITIVITY VALUES

	Gap, X_1 (milli-inches)			
	4		10	
Contact force, F_C (grams)	2		10	
Operate flux, Φ (maxwells)	27		140	
Length, L_2 (cm)	3.65		8.68	
Permeance factor, p	1.76		1.73	
Gap reluctance, $\mathfrak{R}_G$ (cm^{-1})	1.30		0.38	
$\mathfrak{R}_G p L_2$	8.34		5.77	
	Permeability, μ			
	1000	5000	1000	5000
Reluctance per cm, r (cm^{-2})	0.450	0.090	0.086	0.017
αL_2	3.26	1.30	3.18	1.50
$\mathfrak{R} p L_2$	19.0	13.8	17.0	12.5
$\mathfrak{R}$ (cm^{-1})	2.96	2.15	1.11	0.83
Operate ampere-turns	63	46	126	93

length L_1 as half of L_2. There were thus obtained the values of reluctance $\mathfrak{R}$ shown in Table II, and thus finally the operate values, given (in abampere-turns) by $\mathfrak{R}\Phi/4\pi$.

In the range of magnetic reed dimensions and contact and retractile forces illustrated by these computations, αL_2 is small and the reed reluctance minor for μ of the order of 5000. The materials used for these reeds have permeabilities of 5000 and higher at densities of the order of 12,000 gauss, or at about 80 per cent of the saturation density. The permeability falls rapidly as the density approaches saturation, and is of the order of 1000 at about 90 per cent of the saturation density. Hence a design requiring an operating density over 80 per cent that for saturation may have a materially increased reluctance and consequent loss of sensitivity.

VIII. RELEASE RELUCTANCE AND SENSITIVITY

As noted in the discussion of the pull relation, the computed value of flux at which release occurs is given by $\sqrt{8\pi abs_R X}$. Since this is necessarily a lower value of flux than that for operation, the density is lower and the permeability higher than in operation, so that in the release case the reed reluctance is minor, and the preceding equations can be used to estimate the reluctance for the closed gap condition. Since flat metallic pole faces in contact have an effective air gap x_0 of about 2 milli-inches (0.005 cm), $\mathfrak{R}_G$ for this case is given by x_0/ab. Then the abampere-

turns for release are given by $\mathcal{R}\Phi/4\pi$, where $\mathcal{R}$ and Φ are the values applying in release. As previously stated, the actual release flux is variable, and in general less than its computed value, and there is therefore a corresponding variation in release sensitivity.

IX. ALTERNATIVE RELUCTANCE APPROXIMATIONS

The analysis of the two preceding sections is useful in providing an understanding of the field distribution in the reeds and coil and of the way in which the leakage field shunting the gap varies with the gap, the flux density, and the reed and coil dimensions. It is, however, limited to the case of an air return, and does not apply to a configuration such as that shown in Fig. 8, where sleeve members are used to couple the reeds to the return path provided by the cover. In such cases, the reluctance can be approximately estimated in terms of the lumped constants of the magnetic circuit shown in the figure. A similar treatment can be applied to the air return case, giving approximate expressions that are simpler than those developed above.

As indicated in Fig. 8, the coil magnetomotive force is taken as developing an air flux φ_A, which only affects the inductance, and a reed flux Φ, the maximum or total flux Φ of the preceding discussion. The path of Φ comprises the reed reluctance $\mathcal{R}_R$ in series with the sleeve gap reluctance $2\mathcal{R}_S$ and the parallel combination of the gap reluctance $\mathcal{R}_G$ and the shunting leakage reluctance $\mathcal{R}_L$. The path, of course, includes the shielding return members, but these are negligible in reluctance compared with the rest of the path.

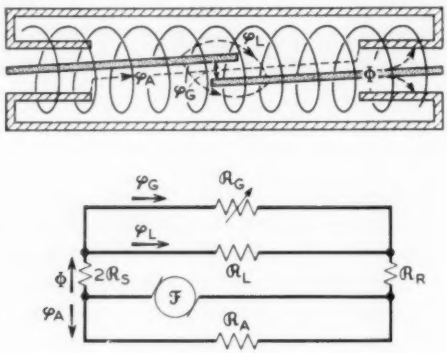


Fig. 8 — Magnetic circuit for reeds coupled to a return path member.

The reed reluctance can be estimated from the reluctance per unit length $r = 1/\mu bh$. The sleeve reluctance can be estimated by an approximation obtained from Fig. 9-8 of Peek and Wagar,[7] from which

$$\mathcal{R}_S = \frac{1}{2\pi L_S} \ln \frac{\pi D}{2(b + h)}, \tag{22}$$

where L_S is the sleeve length, and D its diameter.

As shown in the air return case, the leakage reluctance $\mathcal{R}_L$ varies with the gap reluctance, and cannot readily be expressed in a simple form. An approximate expression for the parallel combination of $\mathcal{R}_G$ and $\mathcal{R}_L$, however, can be derived from the experimentally observed pull relation of (3). The pull given by this relation must equal that given by the general pull equation for a variable reluctance $\mathcal{R}_X$:

$$F = \frac{\Phi^2}{8\pi} \frac{d\mathcal{R}_X}{dx}.$$

If $\mathcal{R}_X$ is the parallel combination of $\mathcal{R}_G$ and $\mathcal{R}_L$, this expression for F/Φ^2 can be equated to that given by (3), and the resulting expression for $d\mathcal{R}_X/dx$ integrated to give the increase in $\mathcal{R}_X$ from its value for $x = 0$. This zero gap value may be taken as x_0/ab, with $x = 0.005$ cm, as in the Table II estimates of $\mathcal{R}_G$. Thus $\mathcal{R}_X$ is given approximately by

$$\mathcal{R}_X = \frac{x_0}{ab} + \frac{1}{kb} \ln \left(1 + k \frac{x}{a}\right), \tag{23}$$

where, as before, k is given by (4).

Then the total reluctance $\mathcal{R}$ for the magnetic circuit of Fig. 8 is given by

$$\mathcal{R} = \mathcal{R}_E + \mathcal{R}_R + \mathcal{R}_X, \tag{24}$$

with $\mathcal{R}_X$ given by (23) and $\mathcal{R}_E$ taken as $2\mathcal{R}_S$, as given by (22). This expression can be used to estimate the reluctance for other relay configurations, using expressions for $\mathcal{R}_E$ corresponding to the return path coupling and configuration. It may be applied as an approximation to the air return path case, with $\mathcal{R}_E$ taken as the air return reluctance. For negligible reed reluctance, this is given by (21) for $\mathcal{R}_G$ and hence Q equal to zero, so that $\mathcal{R}_E$ is given by

$$\mathcal{R}_E = \frac{8}{p(2L_2 - L_1)}. \tag{25}$$

While these expressions for the reluctance are approximate, they are simple in form, and are convenient to use, in conjunction with operate and release flux estimates, in estimating operate and release sensitivity.

X. MOTION TIME

The operate time of magnetic reed contacts is that required for field development and reed motion. These two occur together, since reed motion starts as soon as the field and the resultant pull start to develop. If the power input is high enough for the field to develop rapidly, the reeds may be saturated before any significant motion has occurred. In this case the times of field development t_1 and reed motion t_2 are additive in determining the operate time. The value of t_2 for this case constitutes a lower limit to the operate time, since no shorter motion time is possible than that for full magnetization of the reeds. An expression for this minimum motion time may be obtained by the following approximation.

The pull curve is given by (3), where Φ has its maximum value, which may be taken as about 90 per cent of saturation, as in the capability estimates. This pull exceeds the just-operate pull shown in Fig. 3, and hence always exceeds the stiffness load $s(X - x)$. The area between the pull curve and the load line represents the kinetic energy supplied to the reeds, which may be denoted $2T$, so that T is the kinetic energy supplied to each of the equal reeds. Writing F_0, as before, for the closed gap pull for maximum flux, the kinetic energy is given by

$$2T = \int_0^X \left[\frac{F_0}{1 + Cx} - s(X - x) \right] dx,$$

and hence

$$2T = \frac{F_0}{C} \ln (1 + CX) - \frac{SX^2}{2}. \tag{26}$$

While the acceleration is variable, the motion time is nearly the same as that for uniform acceleration, for which the kinetic energy T for each reed is given by

$$T = 2m \left(\frac{X}{2t_2} \right)^2,$$

where m is the effective mass of each reed, which moves through half the gap X in time t_2. Thus the minimum motion time t_2 is given by:

$$t_2 = X \sqrt{\frac{m}{2T}}, \tag{27}$$

where T is given by (26). To illustrate the magnitude of the quantities involved for mating magnetic reeds, Table III gives the computed values of t_2 for the two cases of Fig. 4 in which the minimum gaps are 4 and 10

TABLE III — MOTION TIME ESTIMATES

X_2 (milli-inches)	X_2 (cm)	F_0 (dynes)	s (dynes/cm)	C (cm^{-1})	$2T$ (ergs)	m (grams)	t_2 (μ seconds)
4.8	0.0122	5,000	246,000	240	10.2	0.0041	245
12.0	0.0305	25,000	491,000	96	128	0.0549	634

milli-inches and the contact forces are 2 and 10 grams respectively. The gap is taken as having its maximum value X_2, and F_0 as corresponding to a density of 13,500 gauss, as before. As the reeds deflect as cantilever beams, the effective mass m of each reed is taken as one quarter of its actual mass, assuming a density of 8.2 grams/cm.[3] The values of C, or k/a, are those applying for the dimensions given in Fig. 4 for these two cases.

These estimates show that for the range of reed and gap dimensions covered in Fig. 4, the motion times lie in a relatively narrow band. These times are small compared with those of most other devices which open and close contacts in a metallic circuit: they are large compared with the switching times of many electronic devices.

XI. OPERATE TIME

To estimate the total operate time, an expression is required for the time of flux development. As shown in Chapter 4 of Peek and Wagar,[7] an adequate approximation can be obtained from the exponential relation for flux development with the constant reluctance (and hence inductance) for the initial open gap condition. For magnetic reeds, eddy current effects are negligible except for flux development so fast that t_1 is negligible compared with t_2. The flux and coil current develop together with a time constant L/R. The time t_1 at which the ampere-turns equal the just operate value $(NI)_0$ is given by

$$\frac{(NI)_0}{NI} = 1 - e^{-Rt_1/L},$$

where NI is the steady-state value of ampere-turns, Hence:

$$t_1 = \frac{L'N^2}{R} \ln \frac{1}{1 - v},$$

where $v = (NI)_0/(NI)$ and L' is the single-turn inductance, or $4\pi/\mathfrak{R}$, where $\mathfrak{R}$ is the open gap reluctance for the average flux linked by the

coil. If both numerator and denominator in this expression are multiplied by I^2, NI is written as $(NI)_0/v$, and t_2 added to t_1 to give the total operate time t, the resulting expression is

$$t = t_2 + \frac{L'(NI)_0^2}{I^2R}\left(\frac{1}{v^2}\ln\frac{1}{1-v}\right). \tag{28}$$

As shown in the text reference cited above, the function of v appearing in brackets has a minimum value of 2.5 for $v = 0.715$, and departs from this minimum by less than 5 per cent for $0.6 < v < 0.8$. Thus for minimum operate time, $(NI)_0/NI$ should lie in this range, and (28) reduces to

$$t = t_2 + \frac{2.5nL''(NI)_0^2}{I^2R}, \tag{29}$$

where nL'' is written for L'. If there are n sealed contacts (pairs of mating reeds) used in a single relay coil, the single-turn inductance is nL'', where L'' is the single turn inductance of one sealed contact.

If the total time t is materially larger than t_2, the two terms of (29) are not strictly additive, but the approximation has been found to give adequate agreement with experimental observations.

In (28), the term $L'(NI)_0^2$ is proportional to $\Phi(NI)_0$ (where Φ is the operate flux) and hence to the field energy required for operation. Thus the electrical energy input for operation, which is proportional to I^2Rt_1, is proportional to this field energy. As previously noted, (28) also shows that t_1 varies with v, or I_0/I, and that the time is a minimum for a given power input if the coil circuit is designed to give I_0/I a value of about 0.7. It also follows from (29) that increasing the steady-state power reduces the operate time until it approaches the lower limit t_2, the minimum time for reed motion.

To use (29) for preliminary estimates requires estimation of the operate ampere-turns $(NI)_0$ and the single-turn inductance nL''. $(NI)_0$ can be estimated by the procedures described in the sections on reluctance estimates. To estimate the single-turn inductance requires estimating the reluctance for the average field linked by the coil. In general, this is a larger reluctance than that for the maximum reed flux, to which the expressions given in preceding sections apply. The maximum flux is limited by reed saturation, and hence controls the capability.

For a tightly coupled magnetic circuit, as in Fig. 8, the maximum and average flux are nearly the same, and the same reluctance expressions apply, except that some allowance may be made for the air field (φ_A of Fig. 8) in estimating the inductance. For the distributed field of

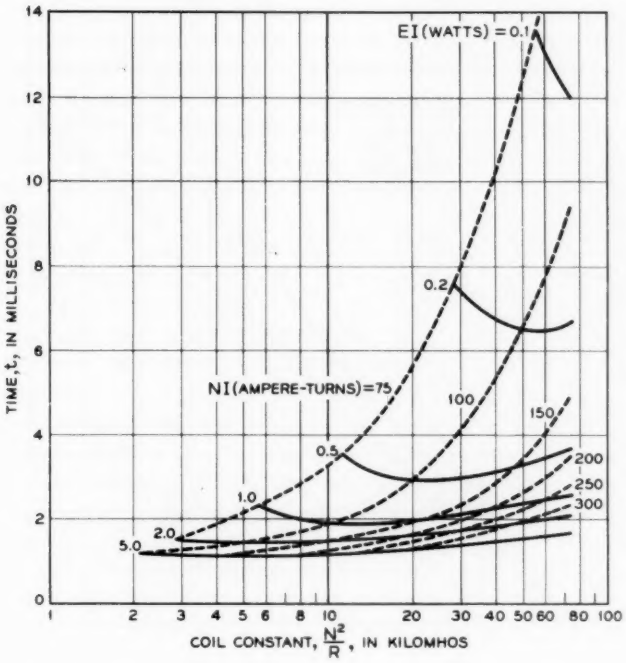


Fig. 9 — Observed operate times for a two-switch reed relay.

an air path return relay, the average flux linked by the coil is materially less than the maximum value. The analysis of this case given above can be used to give expressions for the average flux, but these are complex in form, and are not included here.

The agreement of experimental observations with (28) and (29) is illustrated in Figs. 9 and 10. Fig. 9 shows direct measurements of operate time plotted against the coil circuit constant N^2/R for various values of steady state power input. The dotted curves are for constant values of ampere-turns NI. [$(NI)^2 = I^2R \cdot (N^2/R)$.] The sealed contacts in this relay had a value of $(NI)_0$ of about 100 ampere-turns, for which the minimum times should occur for a value of NI of about 140, which is near that for the observed minima.

The minimum times for this two-switch assembly are plotted against n/I^2R in Fig. 10, together with those for three other cases with one, three and four switches. In agreement with (29) the plots are linear, and have a common intercept t_2, the minimum motion time. These

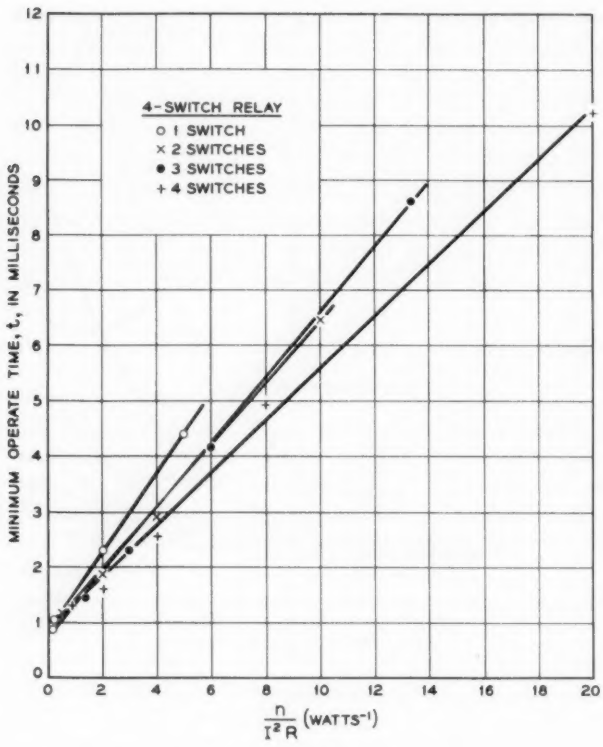


Fig. 10 — Observed minimum operate times for a reed relay.

times were measured to the end of the short chatter interval following initial closure, so that the indicated value of t_2 includes this chatter interval. The slopes of the lines measure the relative value of the single turn inductance per contact L''. The differences in these slopes is ascribed to the considerable air core inductance in this coil, which had space for four switches (sealed contacts). This results in a decrease in the apparent value of L'' as the number of switches increases.

XII. DISCUSSION

The analysis given in this paper is intended for use in development studies of sealed magnetic reed contacts. It can be used to estimate the gap and reed dimensions needed to meet design objectives with respect to contact gap, contact and retractile forces, size, speed and

sensitivity. In development studies, models based on the initial estimates can be made and measured, and the measurement results used in conjunction with the analysis as a guide in further work.

In the performance of mating magnetic reeds, the major controlling quantity is the flux between the reeds. This flux, which determines the pull, is limited by saturation and hence by the reed cross section. This total reed flux comprises the gap flux proper and the leakage flux shunting the gap. As shown in the general analysis of the air return reluctance, this leakage field varies with the gap, reed and coil dimensions, and the relation between the pull and the total reed flux has a corresponding variation. Hence the simple pull equation derived from experiment applies rigorously only for the experimental coil length and return path conditions used in deriving it, and minor deviations occur in applying it to other cases.

Corresponding deviations occur for the reluctance given by the expression derived from the pull equation, and for the other approximate reluctance equations. All these approximations, however, are adequate for preliminary estimates, and may be applied with greater accuracy in detailed studies by using values of the constants that are experimentally determined for the case in question. While the general analysis can be used to estimate the air return reluctance, and the treatment can be extended to give expressions for the pull and the inductance, it is limited to the air return case, and is too complex for convenient application. It serves, however, to show the character and extent of the deviations from the simpler approximations resulting from the distributed character of the leakage field.

As the times of flux development and reed motion depend respectively on the inductance corresponding to the coil flux linkages and on the pull for maximum flux, timing estimates based on the simple approximations given here are also subject to some minor variation. This can be reduced in development studies by using experimentally determined values of the constants.

The expressions given here apply primarily to mating reeds of equal size, but are approximately applicable to unequal reeds provided both are long compared with the gap overlap. In such cases the effective stiffness of the reed combination must be taken as the reciprocal of the sum of the compliances of the two reeds.

XIII. ACKNOWLEDGMENTS

The experimental study described in the section on the pull relation was carried out at Bell Telephone Laboratories by R. Calame and Mrs.

J. Jones, using apparatus designed and constructed by J. L. Agterberg. The author is also indebted to W. B. Ellwood, O. M. Hovgaard and other associates for information used in preparing this article.

REFERENCES

1. Ellwood, W. B., U. S. Patent 2,289,830.
2. Ellwood, W. B., Glass-Enclosed Reed Relay, Elect. Engg., **66,** 1947, p. 1104.
3. Brown, J. T. L. and Pollard, C. E., Mercury Contact Relays, Elect. Engg., **66,** 1947, p. 1106.
4. Hovgaard, O. M., Capability of Sealed Relay Contacts, A.I.E.E. Trans., Pt. I, **75,** 1956, p. 466.
5. Hovgaard, O. M. and Perreault, G. E., Development of Reed Switches and Relays, B.S.T.J., **34,** 1955, p. 309.
6. Hovgaard, O. M. and Fontana, W. J., A Versatile Miniature Switching Capsule, Proc. 1959 Electronic Components Conf., p. 32.
7. Peek, R. L., Jr. and Wagar, H. N., *Switching Relay Design*, D. Van Nostrand Co., New York, 1955.
8. Evershed, S., Permanent Magnets in Theory and Practice, J.I.E.E., **58,** 1919, p. 780.

Lightning Surges in Paired Telephone Cable Facilities

By D. W. BODLE and P. A. GRESH

(Manuscript received October 18, 1960)

The problem of protecting apparatus against lightning surges from connected transmission facilities has become more complex with the use of solid state devices in apparatus design. Consideration of the protection requirements for such apparatus has indicated that existing information concerning the incidence and characteristics of lightning surges is insufficient to develop optimum protection measures. A recently completed field investigation provides additional information in this specific area.

The results of this field investigation and supplemental laboratory surge tests indicate that, in well-shielded underground cable pairs, electrical surges do not exceed approximately 90 volts peak, and that transistorized apparatus capable of withstanding such surge amplitudes needs no further protection. In aerial and buried cable, however, transistorized apparatus requires protection up to the full sparkover potential of 3-mil protector gaps, i.e., to about 600 volts peak. A firm basis for testing and evaluating transistorized apparatus from the lightning surge voltage standpoint is presented.

I. INTRODUCTION

The 3-mil air gap carbon block protector, which has a maximum sparkover value of 600 peak volts, is the basic protection device employed by the Bell System for the protection of communication apparatus against extraneous potentials. Prior to the introduction of transistors and miniaturized circuitry, it was the general practice in apparatus design to provide a withstand level for both metallic* and longitudinal† potentials greater than 600 peak volts so as to coordinate directly with 3-mil protector gaps. This customary design objective of providing an inherent withstand capability exceeding the operating value of standard protector gaps is not always feasible in the case of apparatus employing

* Voltage appearing between the conductors of a pair.
† Voltage appearing between a conductor and ground.

solid state devices. Furthermore, lower voltage protection cannot be attained satisfactorily by simply reducing protector-gap spacing below 3 mils, since excessive protector maintenance would be introduced. To meet the lower voltage requirements of apparatus employing transistors, therefore, it is necessary at present either to modify the circuitry so that surge currents appearing in the more susceptible components are limited in magnitude, or to introduce an additional stage of protection employing semiconductor diodes supplementing the gaps. These protection measures may introduce significant additional expense and, in some cases, produce adverse effects on transmission characteristics.

It became apparent that selection of optimum protection measures to meet the exacting requirements of transistorized apparatus necessitated a more complete knowledge of the incidence and characteristics of lightning surges in the range below the operating level of standard protector gaps. Recognizing this, a field investigation was undertaken to supplement existing information in this area. The results of this recently completed field study and conclusions drawn from analysis of the data are presented in this article.

Since it appeared, at the time the investigation was undertaken, that the present practice of employing 3-mil protector gaps as basic apparatus protection would continue into the foreseeable future, all circuits used for purposes of observation were equipped with such protectors. The area of study therefore was intentionally restricted to surges up to about 600 peak volts as limited by protector operation.

Observations of lightning surges appearing in trunk pairs in aerial and buried cable were recorded by means of automatic cathode ray oscillographs. The plant locations selected were in areas known to experience heavy thunderstorm activity. Surges were also monitored by means of peak amplitude recording devices in urban underground cables. Because of the shielding provided by buildings and buried piping facilities, the exposure of cables to lightning in this situation was relatively low.

Information of engineering value secured includes the probability distribution of voltage magnitudes and the rise and decay time characteristics of surges in the lower voltage range specifically under study. Such data have been used as a basis for selecting waveshapes suitable for laboratory testing of the energy and power handling capabilities of transistorized apparatus.

II. FACILITIES OBSERVED AND MEASURING PROCEDURES

Field data on lightning surge characteristics were obtained from types of telephone plant having two degrees of lightning exposure:

1. Low exposure, typified by underground plant in well-shielded urban areas.

2. High exposure, typified by aerial and buried cables in suburban and rural areas.

2.1 *Low-Exposure Facilities*

The study of surge activity in underground cable was conducted on spare trunk pairs in Baltimore, Maryland; Pontiac, Michigan; and South Orange, New Jersey. Table I gives a brief description of the route and make-up of these facilities.

As indicated in Table I, field observations were made on two types of trunks: those in all-underground cable and those in underground cable with aerial subscriber complements. A total of five trunks for the three locations was monitored for both longitudinal and metallic voltages with gas-tube-type, peak voltage counters. The counters were designed to record surge voltages in three voltage ranges from 90 volts up to the sparkover value of protector blocks.

2.2 *High-Exposure Facilities*

The waveshapes of lightning surges in aerial and buried cable were studied with cathode-ray oscillographs arranged to monitor continuously the pair selected for observation and to record automatically on 16-mm film all surges exceeding 60 volts peak. On each test pair, simultaneous measurements were made of open-circuit longitudinal surge voltages and any resultant metallic voltages appearing across a representative resistive termination. Spare H88-loaded trunk pairs in aerial cable were

TABLE I — DESCRIPTION OF UNDERGROUND TRUNKS MONITORED WITH SURGE COUNTERS

Cable Location	Type of Cable	Circuit Length (miles)
South Orange to West Orange, N. J.	400-pair underground cable with aerial complements	4
South Orange to Summit, N. J.	455-pair 100% underground cable with H88 loading	6
Pontiac to Birmingham, Mich.	Underground cable with aerial complements and H88 loading	8
Baltimore to Pikesville, Md.	900-pair 100% underground cable with H88 loading	8½
Baltimore to Towson, Md.	900-pair underground cable with aerial complements and H88 loading	8

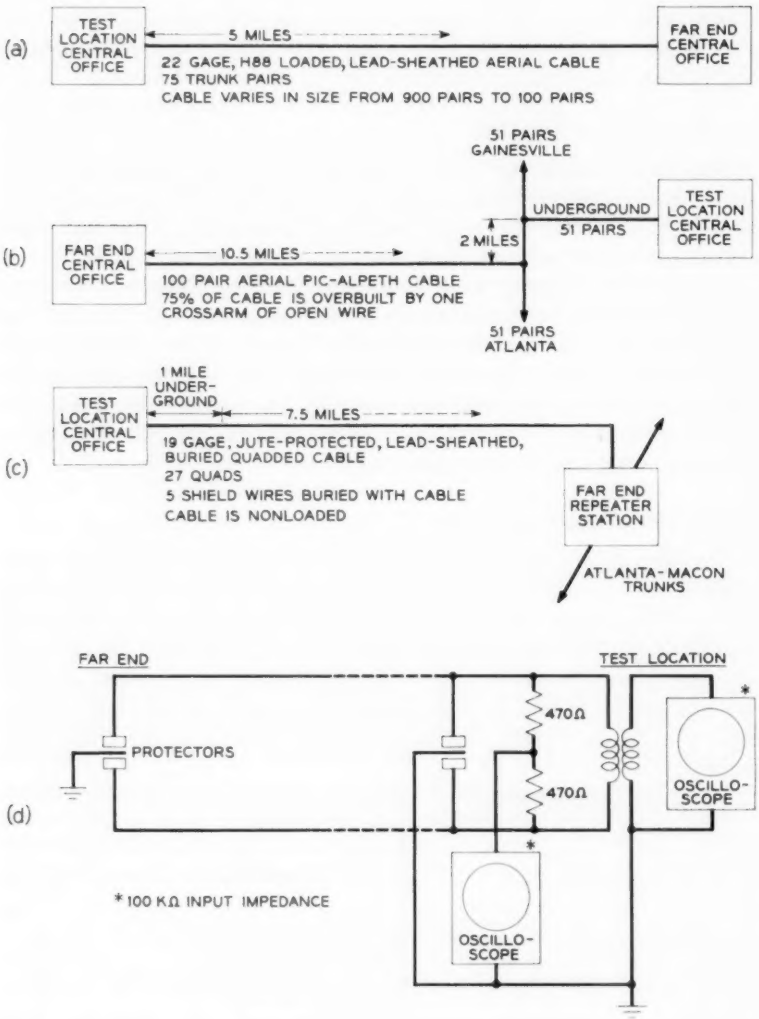


Fig. 1 — Characteristics of test cable at (a) Mt. Freedom, N. J., (b) Buford, Ga., (c) Griffin, Ga.; (d) arrangement of measuring equipment at test locations.

studied in Buford, Georgia, and Mt. Freedom, New Jersey. The buried cable studies were conducted in Griffin, Georgia, on a spare nonloaded trunk. Descriptions of the cable involved and the measuring equipment used at each of these test locations are presented in Fig. 1.

Supplemental laboratory surge tests were also conducted on a one-mile test cable to augment the information on the behavior of underground cables with aerial complements and extensions.

III. RESULTS

The incidence and characteristics of the lightning surges recorded and the resulting protection considerations will be discussed in the order of the degree of plant exposure.

3.1 *Facilities Having Low Lightning Exposure*

Surge characteristics in underground plant were studied in: (a) trunks in all-underground cable, (b) trunks in cable with aerial subscriber complements and (c) trunks extended aerially. The first two situations were studied in the field. The third was investigated subsequently in the laboratory.

The field study of surge activity in well-shielded urban underground plant, covering the first two categories, revealed that in no case did voltages attain the 90-volt triggering value of the lowest stage of the counters. During the five-month observation period, a total of 44 thunderstorm days was reported by the U. S. Weather Bureau for these areas. Several of these storms were known to be quite severe, with their centers located over the test cables. The counters were tested periodically during the study period to ensure proper operation. The lack of surge activity recorded during this study reveals the shielding benefits enjoyed by urban underground plant. In such areas, buildings and buried metallic piping systems divert and dissipate lightning strokes that otherwise might directly involve the telephone cables. Furthermore, duct runs usually contain two or more cables, the sheaths of which are bonded at each manhole. Surge current will, therefore, divide between the cables, and the voltage induced in any one cable will be proportionately reduced.

On the basis of these field studies, it appears that apparatus capable of withstanding surges in the order of 90 to 100 volts peak will not require lightning protection when associated with all-underground cable pairs (trunk or loops), whether such pairs are in an all-underground cable or one with well shielded aerial complements. The significant point

is that this conclusion holds only for well-shielded plant, the shielding being provided by closely spaced buildings, extensive power distribution and buried metallic piping systems, and other telephone cables in the same conduit run.

The question now arises as to the magnitude of surges that may appear in a 100 per cent underground trunk complement through coupling with underground pairs extended aerially in cable having greater exposure to lightning than those employed in the field study.

Information on the following specific points relative to such surge coupling is useful in estimating the secondary exposure likely to be experienced by apparatus connected to the 100 per cent underground trunk pairs:

1. The ratio between an open-circuit longitudinal voltage surge on a disturbing circuit (underground pair associated with an aerial extension) and the resultant voltage appearing in a disturbed circuit (100 per cent underground trunk pair).

2. The resultant magnitude and waveshape of longitudinal current appearing in the disturbed circuit.

To secure this information, supplemental laboratory surge tests were conducted on a 50-pair, one-mile test cable. Longitudinal impulses (approximately 50 by 250 microseconds*) were applied to one or more cable pairs acting as the disturbing circuit. Several cases were investigated: energizing a single pair, then 5 pairs in parallel and finally 25 pairs in parallel. Measurements were made of longitudinal open-circuit voltage and short-circuit current in the disturbing circuit. Measurements were then made of longitudinal open-circuit voltage and short-circuit current in a disturbed pair. Pairs both adjacent and remote from the disturbing circuit were investigated.

The ratio of longitudinal open-circuit voltage appearing in a disturbed pair to the longitudinal open-circuit voltage in the disturbing circuit varied, depending on test conditions, from about 0.47 to 0.85. The lower value was obtained when only a single pair was energized and the larger value when the surge was applied to 25 pairs paralleled at the generator end. Although the magnitude of the open-circuit surge voltage appearing in the disturbed circuit was significantly lower than that in the disturbing circuit, the waveshapes of the two were essentially the same. In the cases investigated it was found that the longitudinal short-circuit current in the disturbed pair was approximately 3 milliamperes per volt appearing in the disturbing circuit. The short-circuit current in the

* That is, 50 microseconds rise time to crest and 250 microseconds from origin to point where wave has decayed to half of crest value.

disturbed pair assumed the shape of a square pulse about 20 microseconds in duration. From a protection standpoint, this radical reduction in the duration of the induced current wave in the disturbed circuit is probably the most significant bit of information secured in these laboratory tests. Semiconductor components, when exposed to lightning surges, usually fail as a result of overheating of their junction or junctions — the heating effect being related to the magnitude and duration of the junction current. Therefore, the possibility of failure from longitudinal current of semiconductor components in apparatus associated with a disturbed pair is much reduced because of the relatively short duration of the coupled surge.

It is of further interest to note the effect of grounding the disturbing circuit at the far end, such as would occur with protector operation. This condition was investigated by grounding one of the conductors constituting the disturbing circuit, and it was found that the open-circuit longitudinal voltage on the disturbed pair was reduced approximately 50 per cent and the short circuit longitudinal current 30 per cent below the values that would obtain if the conductor of the disturbing circuit had not been grounded. This indicates the order of beneficial shielding enjoyed by the disturbed circuit when protector blocks operate on the disturbing circuit.

Additional laboratory surge studies were made employing a one-mile test cable to determine the protection requirements for apparatus connected to underground pairs extended aerially. These tests revealed that surges having typical waveshapes will propagate longitudinally on a cable pair for one mile with little attenuation. Therefore, underground cable pairs extended aerially or in buried plant should be considered as exposed to lightning, unless protection is applied at the underground junction to limit surges from the exposed extensions.

3.2 *Facilities Having High Lightning Exposure*

The lightning exposure of buried and aerial cable is sufficiently severe to require supplementary protection for transistorized apparatus associated with this type of plant. The most useful way of defining protection requirements in this situation is in terms of a simulated lightning surge which such apparatus must withstand in laboratory tests. Simulated surges, of course, must be based on surge conditions in the field. Consequently, derivation of suitable test surges requires knowledge of the distribution of waveshapes and peak voltages of lightning surges appearing in the plant.

3.3 *Longitudinal Surges*

The data on longitudinal surges in aerial cable at Buford covered a period of six months, during which time six thunderstorm days occurred and 103 oscillograms were obtained. Additional data on aerial cable were secured at Mt. Freedom, with 105 oscillograms being recorded during three thunderstorm days. The buried cable studies conducted at Griffin covered a period of six months, during which 36 thunderstorm days occurred and 1120 oscillograms of longitudinal surges were obtained. Since these cables involved very little unexposed plant, the surge magnitudes and waveshapes recorded at the central offices should also be reasonably representative of surge conditions along the cable route.

Comparison of the recorded waveshapes of longitudinal surges induced in aerial and buried cables indicates that the two types of plant do not differ significantly in their response to lightning surges. The data also confirm that load coils have little or no effect on the waveshape of longitudinal surges. This observation is based on the similarity of the surges recorded on H88-loaded aerial cable and those recorded on non-loaded buried cable. These surges were found to be essentially impulses, as exemplified in Fig. 2. Longitudinal impulses can conveniently be characterized on the conventional basis of crest magnitude, time to crest and time from origin to the point at which the wave has decayed to one-half of crest value.

Both the rise time to crest and decay time to half-crest value of lightning surges observed in cable exhibited log-normal distributions.

The peak voltages of surges induced in cable pairs were found to follow an exponential distribution similar to the lightning stroke currents that produce these voltages.

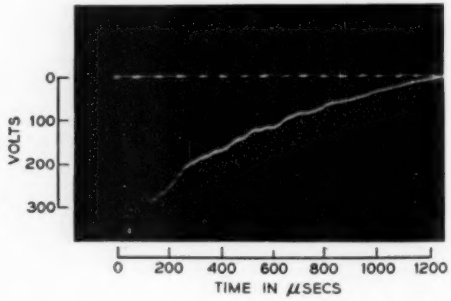


Fig. 2 — Representative longitudinal lightning surge in cable plant.

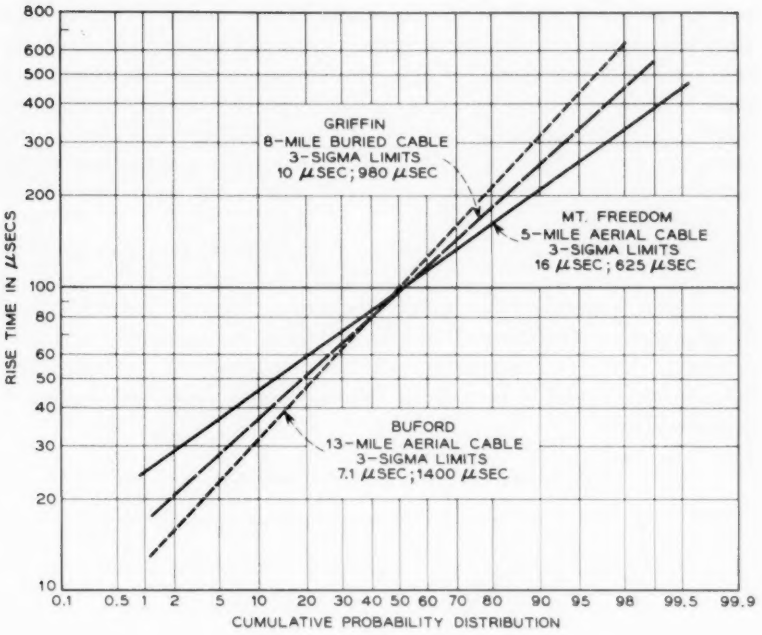


Fig. 3 — Distribution of rise times.

3.3.1 *Rise Time Characteristics*

Distribution of surge rise times for the three test locations are presented in Fig. 3. The median rise time measured at each location was approximately 100 microseconds, but the dispersion about the median value varied widely among locations, probably due to the difference in cable lengths involved. As a surge propagates along a cable pair, there is some modification in waveshape. Consequently, slightly greater dispersion should be expected in longer cables. This is borne out by the measurements plotted in Fig. 3.

3.3.2 *Peak Voltage Characteristics*

The peak voltages recorded in aerial and buried cable in the range below protector block operation (voltages less than 400 volts) were found to be exponentially distributed. This distribution provides a basis for determining the probability of any given surge exceeding a particular value. The derivations of these probability functions for aerial and

buried cable are given in the Appendix. These probability functions, used in conjunction with the average number of surges induced in the test cables per thunderstorm day, provide an order-of-magnitude estimate of the number of surges per thunderstorm day exceeding any given amplitude. The similarity between the peak voltage distributions for aerial and buried cable makes it feasible to develop a single plot of the estimated number of surges per thunderstorm day as a function of voltage.

Fig. 4 presents both the distribution that would be expected if no protector blocks were associated with the test pair and the modified distribution reflecting the operating characteristics of standard 3-mil air gap carbon protectors. This information is useful in the design of reliability tests for apparatus vulnerable to repeated low-amplitude voltage surges, and is used in the selection of suitable test surges, as discussed later.

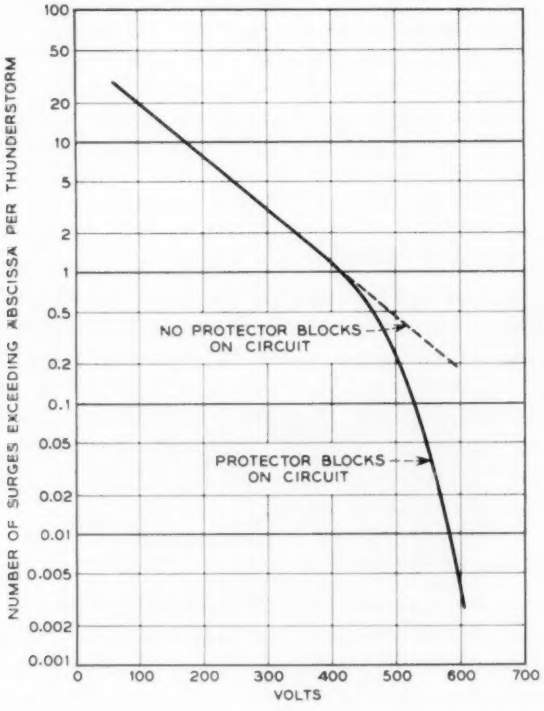


Fig. 4 — Surge voltage distribution on buried and aerial cable.

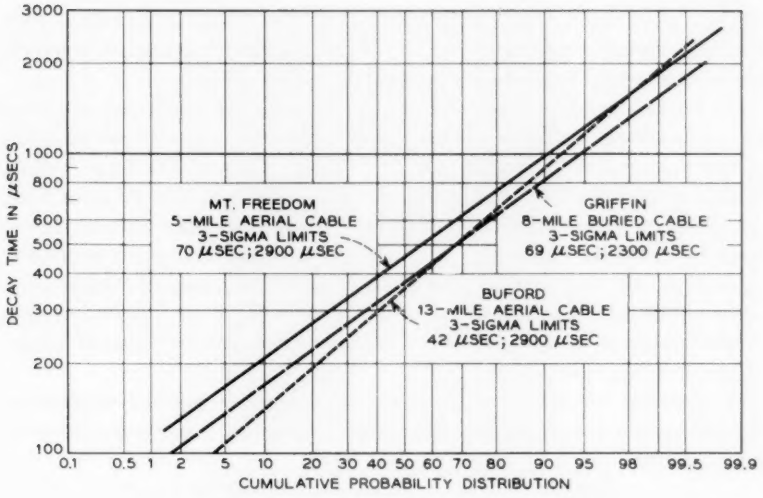


Fig. 5 — Distribution of decay times.

3.3.3 *Decay Time Characteristics*

The distributions of decay times of lightning surges appearing in cable plant are presented in Fig. 5. The median values and dispersions differed slightly among the three test locations. Median values ranged from 350 to 450 microseconds. Three-sigma limits ranged between 40 to 70 microseconds on the low side and between 2300 and 2900 microseconds on the high side. The variation of median values is possibly due to differences in the relative exposure of the test cables.

In Buford and Griffin, the cables traverse relatively level terrain and, therefore, should have fairly uniform exposure along their lengths. Close correlation is noted between the median decay time values for these cables. In Mt. Freedom, however, the cable is placed on hilly terrain, which results in a higher exposure for some sections of the cable. With the majority of the surges being induced into one section of the cable, a shift in the median value results.

3.4 *Metallic Surges*

Metallic surges were recorded on the same trunk cable pairs used at Buford and Griffin to record longitudinal surges. Measurements were made across a 940-ohm resistive termination. Simultaneous operation

of the automatic recording oscillographs observing longitudinal and metallic voltages was arranged in order to permit correlation between each longitudinal surge and any resultant metallic surge.

In the absence of conductor insulation breakdown or protector operation, the circuit balance of cable pairs is sufficiently good that metallic surge voltages should be much lower than longitudinal surge levels, which is confirmed by the measurements secured in this study. Of all longitudinal surges exceeding 60 volts* only about 10 per cent produced metallic voltages exceeding 10 volts peak. However, when protector operation occurred, metallic surge potentials of significant magnitudes were produced. An approximate breakdown of those metallic surges which exceeded 10 volts peak is presented below on the basis of waveshape and magnitude:

1. Twenty per cent were low-amplitude, high-frequency oscillations having maximum peaks of about 35 volts and frequency components ranging from 10 kc to approximately 50 kc.

2. Twenty-five per cent were impulses ranging in amplitude up to about 60 volts peak, which were probably caused by protector operation on adjacent pairs.

3. Fifteen per cent ranged in amplitude from 120 to 200 volts peak. The oscillograms of the associated longitudinal surges definitely indicated protector operation on the test pair although the surge amplitudes were considerably lower than the level normally required for protector block operation. The protector blocks on the test pair were changed periodically during the study period, but these low-voltage operations (16 in all) occurred with the same set of blocks. However, visual inspection of these protector blocks by local personnel did not reveal anything unusual in their appearance.

4. Forty per cent of the metallic surges exceeded 350 volts peak. These surges were all associated with protector operation on the test pair.

Examination of the oscillograms of these high-amplitude metallic surges and the corresponding longitudinal surges illustrates some interesting aspects of protector block operation. Fig. 6(a) shows one type of single-block operation where the discharge is continuous for the duration of the surge. Fig. 6(b) illustrates another type of single-block operation in which clearing and restriking of the arc discharge occurs. This situation is the result of circuit regulation when the longitudinal surge potential is just sufficient to initiate gap sparkover.

* Threshold value of automatic recording oscillographs measuring longitudinal surges.

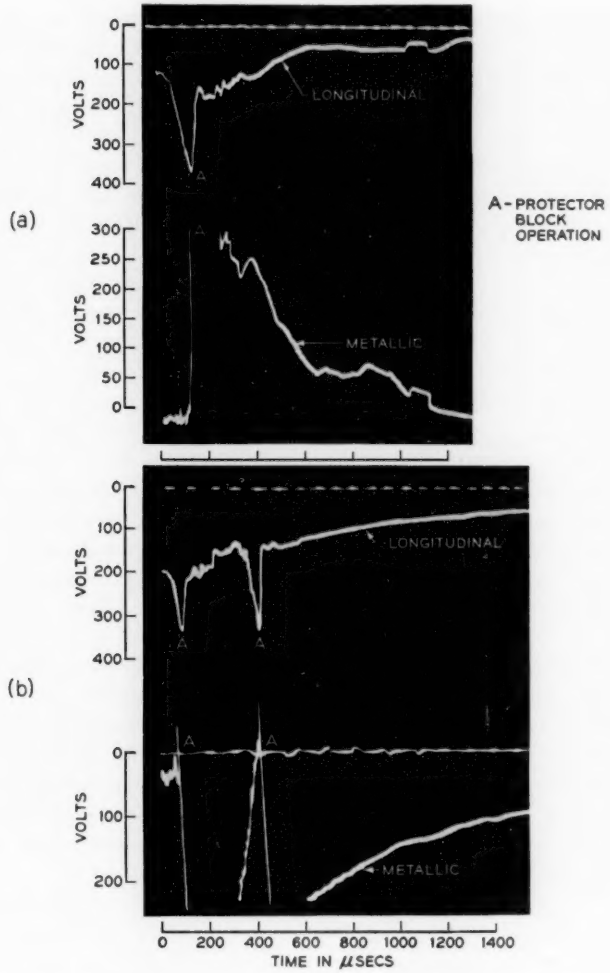


Fig. 6 — Lightning impulse voltages on a nonloaded buried cable pair: (a) with protector operation on one conductor; (b) with multiple protector block operation on one conductor.

A brief explanation of surge voltage relationships during single-block operation as indicated by the oscillograms follows:

At point A [Fig. 6(a)], the gap associated with one conductor operated and remained operated for the duration of the surge. The longitudinal voltage dropped to about one-half of peak value as the oscilloscope is

connected to the midpoint of a 940-ohm termination (see Fig. 1).* The metallic surge shows the true voltage unbalance, which is the difference between the arc discharge voltage on one conductor (approximately 50 volts) and the longitudinal surge voltage on the other conductor.

At the first point A [Fig. 6(b)], the gap associated with one conductor operated and remained operated for approximately 350 microseconds. The metallic surge began at point A and continued for the same duration. At the second point A the protector gap "cleared" and then operated again. When it cleared, the longitudinal scope recorded the true longitudinal voltage* and the metallic voltage dropped to zero. When the block restruck, the longitudinal scope again read one-half true longitudinal voltage and once more there was a metallic voltage.

Metallic surges resulting from operation of both blocks on a pair are shown in Figs. 7(a) and 7(b). In the first case, one block operated initially, then cleared, and then the other block operated. This sequence of operation produced a metallic surge having impulses of both polarities. In the second case, both blocks operated, but unbalances were produced by nonsimultaneous clearing and restriking of the two gaps.

Further explanation of surge voltage relationships resulting from a longitudinal surge of sufficient magnitude to operate both blocks is as follows:

At the first point A [Fig. 7(a)], the protector on one conductor operated, in this case, the one on the "tip" conductor. The recorded longitudinal voltage then dropped to one-half of the actual value and the metallic surge began. At the second point A, the protector on the "ring" conductor operated and the longitudinal surge voltage dropped to essentially zero. At the third point A, the protector on the "tip" conductor cleared. This is evidenced by the reappearance of metallic voltage of reversed polarity.

Fig. 7(b) illustrates a phenomenon that commonly occurs during the operation of protector gaps on telephone circuits. The discharge is not a continuous process but is punctuated by a random restriking of the arc discharge. In this case there was sufficient longitudinal potential to operate both blocks, but apparently some differences in electrode conditions caused nonsimultaneous clearing and restriking of the arc discharge which produced metallic potentials. It is only during the brief

* In effect, the longitudinal scope always reads the true open circuit voltage until one block operates; then it reads one-half the true longitudinal voltage. The metallic oscilloscope always reads the total voltage difference between the two conductors of the pairs.

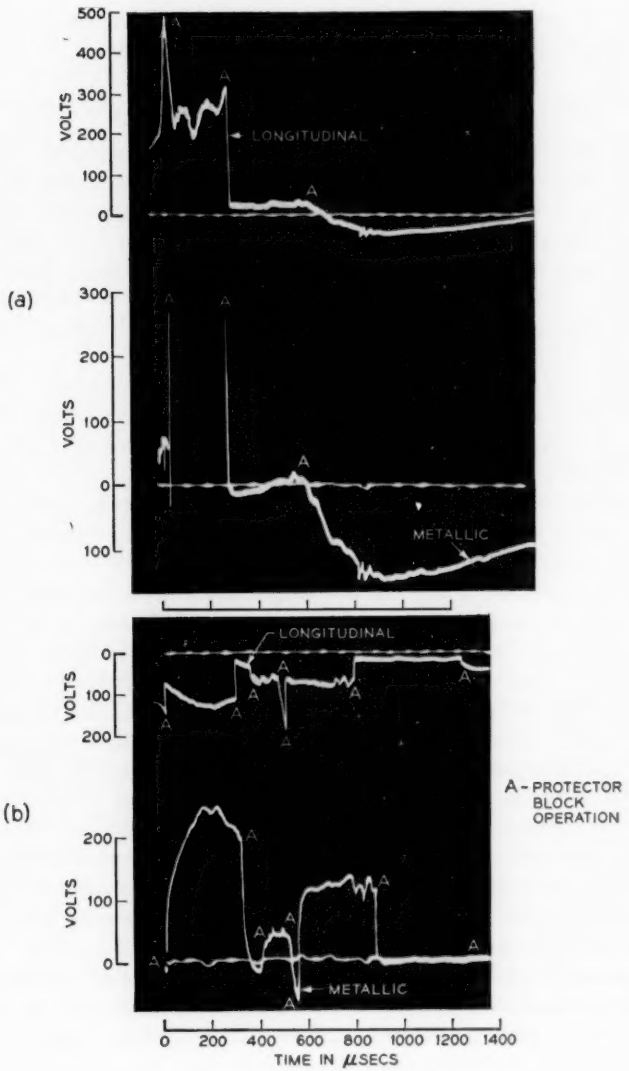


Fig. 7 — Lightning impulse voltages on a nonloaded buried cable pair: (a) with protector block operation first on one conductor, then on the other; (b) with multiple striking and restriking of protector blocks.

periods when the two gaps are either both discharging or not discharging that metallic potentials are reduced to negligible values.

IV. SELECTION OF SUITABLE WAVEFORMS FOR LABORATORY TESTING

The similarity of longitudinal waveforms observed in buried and aerial plant under loaded and nonloaded conditions makes it practicable to employ common waveforms for laboratory testing of apparatus intended for use with all of these types of plant. In the past, a 10 by 600-microsecond surge was selected on the basis of limited field data. The supplemental surge data obtained in this study makes it possible to select waveshapes for laboratory testing which more closely simulate surges produced in the telephone plant by natural lightning.

Analysis of the recorded data obtained during this study provided the distributions of rise times, peak voltages and decay times presented in Figs. 3, 4 and 5. Using these distributions as a basis, it is possible to select suitable laboratory test surges.

Since the severity of a surge is dependent on its peak voltage and its decay time, it is necessary to establish whether decay time is independent of peak voltage before computing the joint probability of occurrence of a surge with a given amplitude and decay time. Accordingly, new decay time distributions were developed from the recorded data for two voltage ranges; voltages below 225 volts and voltages above 225 volts. Fig. 8 presents the results of this analysis. It will be observed that there is a slight correlation between the decay time and voltage. This is the result of the manner in which the surge current in the sheath induces voltage onto the cable pairs. The capacitive coupling between the sheath and the core results in an integration of the sheath current. Thus, surges of longer duration will tend to produce higher voltages on the cable pairs. In determining the joint probability of exceeding a given amplitude and a given decay time, this correlation, as indicated by the field data, should be included by using the probability distribution of decay times associated with higher voltages (upper curve in Fig. 8).

The parameters of laboratory test surges should be so selected as to evaluate apparatus properly for its dielectric strength and its energy and power handling capabilities. The effect of these factors on the parameters of test surges is discussed below.

Surges suitable for test purposes should have a peak amplitude of at least 600 volts to provide a minimum test of apparatus dielectric strength, since 3-mil protector gaps associated with apparatus assure protection only against surges in excess of this value. Furthermore, when the thermal time constants of vulnerable components are small, power is

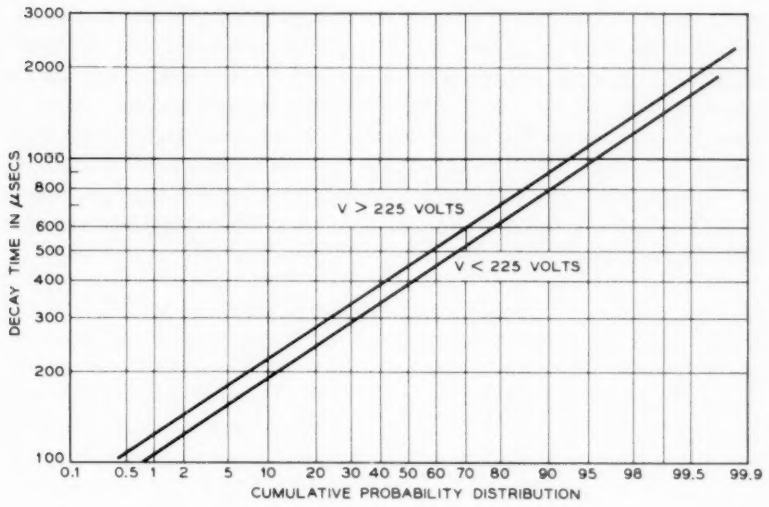


Fig. 8 — Distribution of decay times.

more detrimental than energy content and, given two surges of equal content, more power will be delivered by the surge with the higher amplitude.

The energy content of a surge is dependent on its peak voltage and waveshape (i.e., rise time and decay time). Test surges should have short rise times for two reasons. A short rise time will provide a more severe test of inductive circuit elements and, for a given decay time, the shorter the rise the higher will be the total energy of the surge. Accordingly, a 10-microsecond rise time has been selected as it is approximately the lower 3-sigma limit of the rise times recorded in the field.

In those instances where the energy-handling capability of apparatus is the controlling factor, the reliability of a surge testing program will depend on the degree of assurance that the energy content of the test surge will not be exceeded in the field. The total energy of an impulse with a short rise time is proportional to the decay time and the square of the voltage. The energy content of any arbitrarily assumed test surge can be exceeded in the field in two ways: surges of lower amplitudes but appreciably longer decay times, or surges of higher amplitude and only somewhat shorter decay times. This second classification can be eliminated, however, by selecting test surges having peak amplitudes of 600 volts or greater, as the standard 3-mil air gap protectors associated with equipment will not permit surges in excess of 600 peak volts. To deter-

mine the probability of exceeding the energy of a particular test surge, the joint probability of two factors must be calculated: (a) the probability of obtaining a particular voltage and (b) the probability that a surge of this voltage has a sufficiently long decay time such that its total energy exceeds the energy content of the test surge. These joint probabilities must be summed for all surges with realizable combinations of voltages and decay times which exceed the energy of the test surge.

Mathematically the procedure is as follows: The energy content of a lightning impulse* is:

$$\text{Energy } (E) = \int [f(t)]^2 \, dt,$$

where

$$\begin{aligned} f(t) &= 0 && \text{for} \quad t < 0 \\ &= \frac{vt}{t_0} && \text{for} \quad 0 < t < t_c \\ &= ve^{-b(t-t_0)} && \text{for} \quad t > t_0, \end{aligned}$$

v = peak amplitude of impulses,

$$b = \frac{0.69}{\text{decay time to half value}} = \frac{0.69}{d},$$

t_0 = time to crest amplitude.

Therefore,

$$E = \int_0^{t_0} \left(\frac{v}{t_0}\right)^2 t^2 \, dt + \int_{t_0}^{\infty} v^2 e^{-2b(t-t_0)}.$$

However, due to the small percentage of the total energy produced during the rise time, the surge energy can be approximated by:

$$E = \int_0^{\infty} v^2 e^{-2bt} \, dt = \frac{v^2 e^{-2bt}}{-2b}\Big|_0^{\infty} = \frac{v^2}{2b} = \frac{v^2 d}{1.38}.$$

Assuming a test surge of amplitude V_0 and decay time d_0, the probability $[P]$ that a surge of somewhat lower amplitude V_1 will have greater energy content is determined as follows:

$$[P] = [P(V_1)][P(d > d_1)].$$

* Impulse = rapid rise to crest amplitude followed by a slow exponential fall.

But, to insure that surge V_1 has an energy content equal to test surge V_0,

$$\frac{V_1^2 d_1}{1.38} = \frac{V_0^2 d_0}{1.38}.$$

Therefore,

$$d_1 = \left(\frac{V_0}{V_1}\right)^2 d_0.$$

To determine the total probability of surges encountered in the field having greater energy content than the test surge requires the summation of $[P]$ for all possible values of V. These probabilities need be summed for voltages only above 400 volts, since lower voltages would require associated decay times far longer than observed or expected in cable plant. This follows for two reasons: In order for a small amplitude surge to have an energy content equal to that of the test surge, its decay time must increase as the square of the voltage ratio of these surges. This would require existence of decay times much longer than 2,000 microseconds, a condition generally contrary to all test observations, and contrary to the observed correlation between peak voltages and decay times which indicated that lower voltage surges have smaller decay times.

Therefore, the total probability (P_T) of exceeding the energy of a particular test surge is:

$$(P_T) = \sum_{v=400v}^{v=600v} [P(\Delta v_1)][P(d \geqq d_1)],$$

where Δv = specified voltage increments

$$d_1 = \frac{V_0^2 d_0}{V_1^2}.$$

Thus we have:

$$\begin{aligned}
[P(400 - 410)][P(d \geqq d_1)] &= A_1 \\
\cdot \quad \cdot \quad \cdot &\quad \cdot \quad \cdot \\
\cdot \quad \cdot \quad \cdot &\quad \cdot \quad \cdot \\
[P(590 - 600)][P(d \geqq d_x)] &= A_x \\
(P_T) &= \sum.
\end{aligned}$$

The probability of obtaining a voltage in the range Δv may be determined from Fig. 4. For example:

$$P(520 < v < 530) = P(v > 520) - P(v > 530) = 0.12 - 0.09 = 0.03.$$

The probability of obtaining a decay time greater than d_1 may be read directly from the upper curve of Fig. 8.

This type of calculation must be repeated for each test surge desired. In this study, test surges with peak amplitudes of 600, 700, 800 volts and a number of different decay times were examined. By providing several waveshapes with equivalent energy content: (a) a better indication may be obtained as to the adequacy of the thermal time constants of vulnerable apparatus components, (b) more latitude is provided for determining the dielectric strength margin of apparatus, and (c) flexibility is provided in the selection of laboratory surge forming circuit constants. The above calculation indicates the probability, per thunderstorm, of exceeding the energy content of a particular test surge. For engineering purposes, however, it is more practical to develop curves of apparatus energy-handling capabilities versus trouble expectancy in years rather than in thunderstorm days. Reference to isoceraunic charts indicates thunderstorm incidences varying from 5 to 90 thunderstorms per year for various sections of the country, with the higher incidences in the Southeast. However, even when a thunderstorm is reported in the general area of a particular cable it is not necessarily close enough to induce surges having magnitudes of the order under discussion. It is felt that an average of only 25 thunderstorm days per year in the higher storm incidence areas are likely to produce significant surges in the cable plant. This factor, together with the calculated probability per thunderstorm of exceeding various energy levels, provides the basis for the curves presented in Fig. 9. These curves indicate the probable surge trouble rate of apparatus tested with surges having parameters that will just cause failure.

The curves in Fig. 9 provide a means of determining the probable lightning surge-handling capabilities of apparatus in two ways. First, where the acceptable lightning trouble rate has been established by system requirements, the parameters of the appropriate test surge may be read from the curves. A second approach may be employed in the case where it is desired to determine the probable trouble rate with regard to a specific piece of apparatus. The procedure would be to establish, by tests, the withstand level of the apparatus in question by employing surge waveshapes selected from the curves. After determining the withstand point, the corresponding estimated trouble rate may be read from the curves.

The waveshapes presented in Fig. 9 will concurrently test apparatus for its dielectric strength, its energy-handling capabilities and its ability to dissipate power. Although these test surges were developed on the

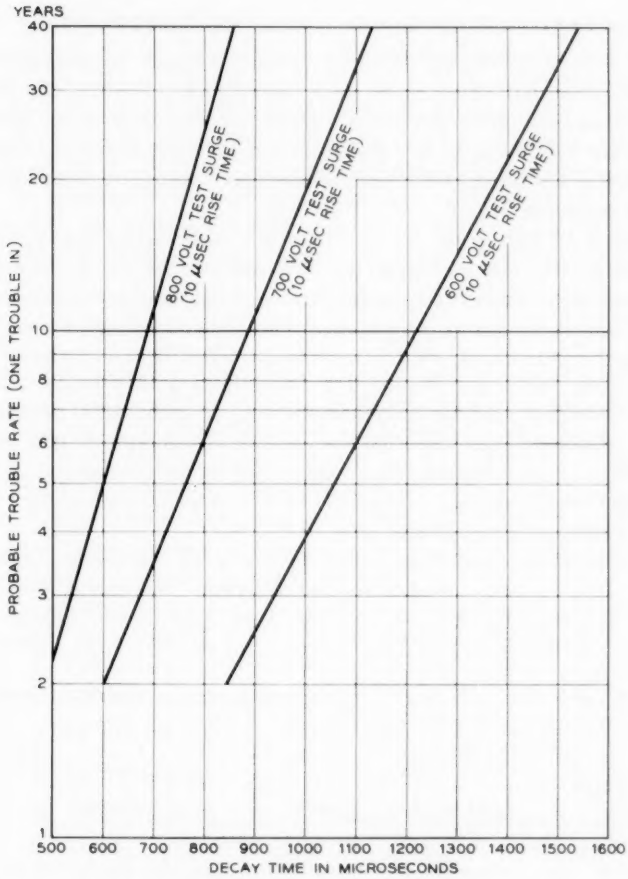


Fig. 9 — Relationship of probable trouble rates to test surges having parameters that will just cause equipment failure.

basis of longitudinal surge data, it is felt that they will provide a reasonable test of the surge-handling capabilities of apparatus, both longitudinally and metallically. Since the total energy contained in a metallic surge must be somewhat less than the total energy of the associated longitudinal surge, the recommended test surges, when applied metallically, will provide an added safety factor. However, this additional safety factor cannot be specifically evaluated from the limited metallic surge data obtained in the study.

V. ENERGY SPECTRUM OF LIGHTNING SURGES IN CABLE PLANT

The main detrimental effect of lightning surges on semiconductor devices is excessive heating of their junction or junctions, the heating effect being related to the energy content of the surge. Since semiconductor devices are used in circuits having different frequency response ranges, it is desirable to determine the energy versus frequency distribution of lightning surges.

Analysis of the energy versus frequency distribution was performed on a 25 by 160-microsecond surge, as it represented one of the shortest duration surges observed on cable plant in this study, and therefore contained higher frequency components. The analysis indicates that 85 per cent of the total energy of this wave is contained in the frequency band up to 3,500 cycles per second. For comparison with a longer duration surge, a similar analysis was performed on a 10 by 1000-microsecond surge, which indicated that 90 per cent of the energy of this surge is contained in the frequency band up to 660 cycles per second. Plots of the cumulative per cent energy as a function of frequency for both of these surges are given in Fig. 10.

In view of the limited frequency spectrum of the energy content of longitudinal surges, it is desirable to determine the energy versus frequency distribution of those metallic surges which are oscillatory. The oscillograms of metallic surges showed the shortest time interval be-

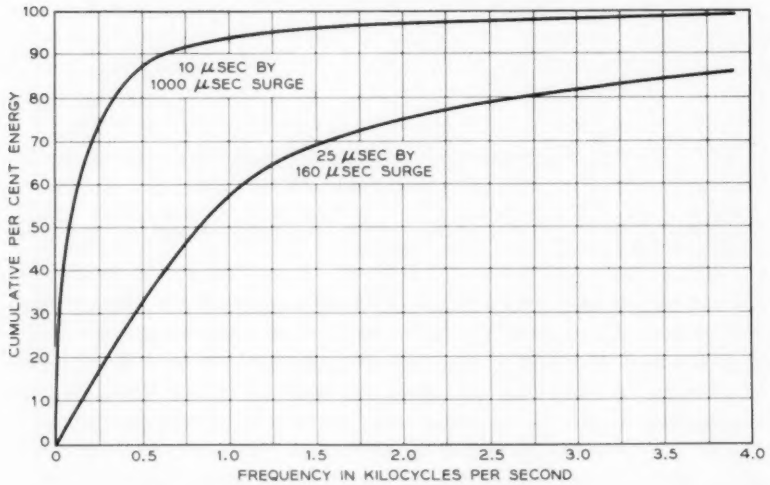


Fig. 10 — Cumulative distribution of energy vs. frequency for two sample surges.

tween polarity reversals to be 80 microseconds. An energy-frequency spectrum analysis for such a wave indicates that 90 per cent of the energy is contained in frequencies below 7 kc. It therefore appears that most of the energy of metallic surges appearing across a resistive termination is likely to be in the frequency band below 7 kc. This does not, however, preclude the need for lightning protection of apparatus operating at carrier frequencies. If reactive components are present in the metallic termination, the resulting metallic surges may have considerable energy in the higher-frequency bands because of spurious oscillations.

VI. SUMMARY AND CONCLUSIONS

Lightning surges were recorded in trunk pairs in aerial and buried cable at several locations known to experience heavy thunderstorm activity. Surges were also monitored in trunks in well-shielded underground cables in urban areas. Observations included measurement of longitudinal surge voltages (from conductor to sheath or ground) and metallic surge voltages (between conductors of a pair). Supplemental laboratory measurements of surge characteristics in simulated underground plant were made using test cable.

In the underground cables monitored, no surges appeared in excess of 90 volts (the minimum sensitivity of measuring equipment) although a total of 44 thunderstorm days occurred during the observation period at the three test locations. Based on this field experience and supplemental measurements made on test cable, it was concluded that apparatus capable of withstanding surges up to 90 volts peak should not require lightning protection if connected to well-shielded all-underground cable pairs. This category includes underground trunks in cables with well-shielded, aerial subscriber complements such as block cable. However, underground pairs extended aerially, or in buried plant, should be considered as exposed to lightning, unless protection is applied at the underground junction to suppress surges from the exposed extensions.

In the aerial and buried cable plant studied, about 1400 surges were recorded, ranging in peak amplitude from 60 volts (minimum sensitivity of equipment) up to 450 volts, the value at which carbon block protectors operated. About 90 per cent of the recorded surges were longitudinal; the remainder were metallic. Analysis of this data helped establish the relationship between the parameters of specific test surges and the probable lightning trouble rate of exposed transistorized apparatus. This information will facilitate appropriate laboratory testing of apparatus for specific levels of reliability.

APPENDIX

With the recording technique used in this study, only those surges in the range above 60 volts and less than the operating value of carbon block protectors were registered. The 3-mil gap, carbon protector blocks associated with "exposed" cable plant limit maximum voltages to a nominal value of 500 volts. Due to manufacturing and field duty variations, the operating value of these carbon blocks may vary from approximately 400 to 600 volts. Protector block operation, therefore, affected the distribution of peak voltages above 400 volts. The recorded data established the distribution of surge voltages between the limits of 60 volts and approximately 400 volts. This distribution was extended to 600 volts by considering the effects of block operation in the 400 to 600 volt range as discussed later.

In the aerial cable plant at Buford, Georgia, a total of 103 surges was recorded during the six-month study period. In the buried cable plant at Griffin, Georgia, a total of 1120 surges was recorded for the same period of time. Histograms of the distribution of these surges as a function of voltage are presented in Figs. 11 and 12 for the two types of

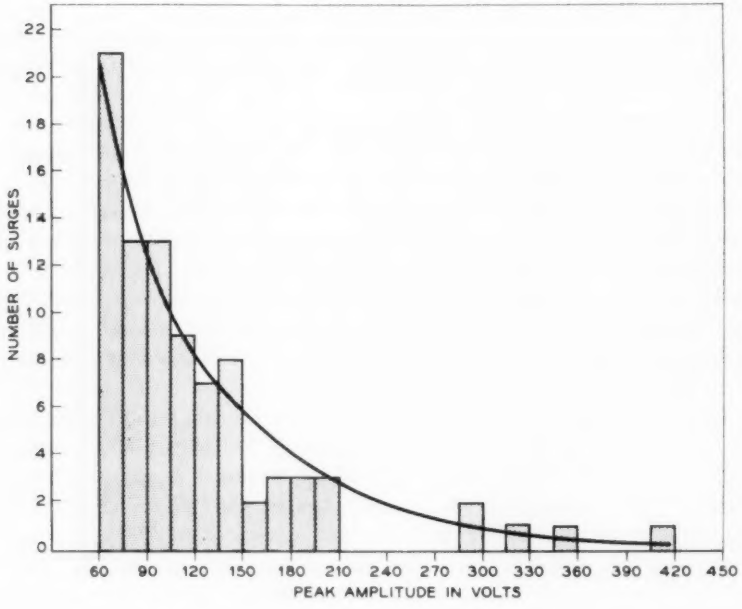


Fig. 11 — Amplitude distribution on aerial cable.

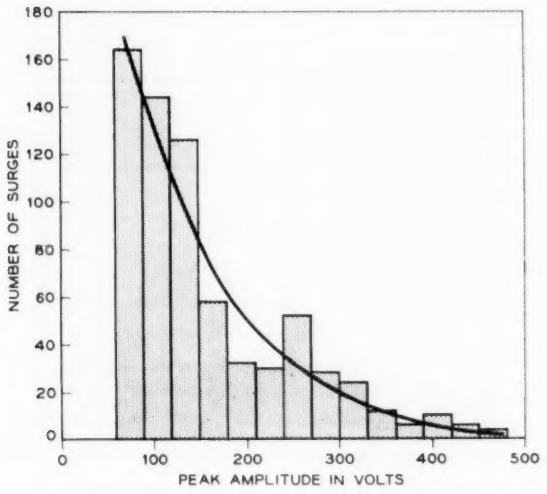


Fig. 12 — Amplitude distribution on buried cable.

plant. These distributions are of an exponential form and can be closely approximated by

$$y = ce^{-aV},$$

where

y = number of surges,
V = peak voltage of the surge.

Smoothing of the raw data (fitting of an exponential by the method of least squares) resulted in the plots on Fig. 13. The constants for the exponential distribution on aerial cable are $a = 0.012$ and $c = 40$. The constants for the buried cable are $a = 0.0094$ and $c = 330$.

The large difference in the calculated values of c for the two types of plant merely reflect the difference in the sample sizes and will not affect their probability distributions. It will be noted that the values calculated for the constant term a for the two types of plant varies over 20 per cent, but it is recommended that the smaller value, calculated for buried cable, be used for both types of cable installation for two reasons. First of all, much more data were obtained in the buried cable study, which permitted a more accurate determination of the shape of the exponential distribution. Also, since our objective is to provide suitable surge voltages for laboratory testing of apparatus, it is safer to use the

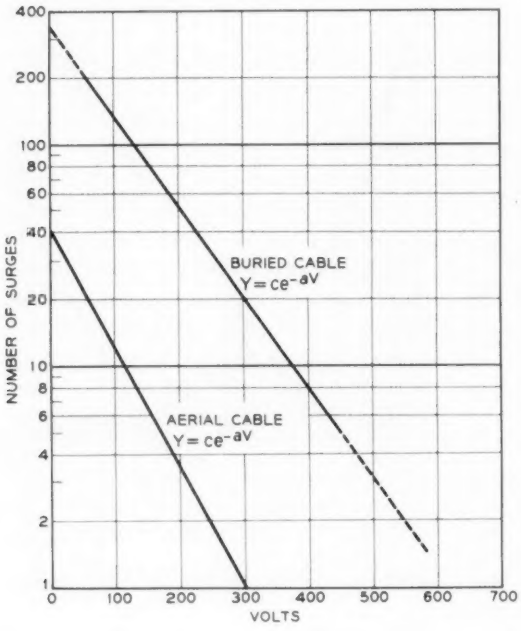


Fig. 13 — Amplitude distribution of surges on aerial and buried cable.

distribution with the smaller value of a which predicts a greater percentage of higher voltage surges.

Having determined the distribution of surge voltages, it is possible to estimate the probability that V is less than or equal to some value T:

$$P(V \leqq T) = \int_0^T f(x)\,dx = \int_0^T Ce^{-aV}\,dV$$

$$= -\frac{C}{a}e^{-aV}\Big|_0^T = \frac{C}{a}(1 - e^{-aT}).$$

This equation must be normalized, since

$$P(V > 0) = 1.$$

Therefore,

$$P(V \leqq T) = 1 - e^{-aT}.$$

Using the previously established value of a, the probability distribution of surge voltages was then plotted (Fig. 14). This, however, gives

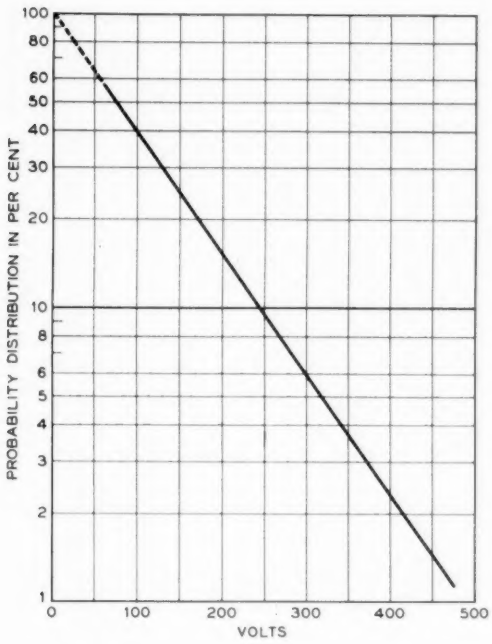


Fig. 14 — Probability distribution of all surges in buried and aerial cable.

the distribution of *all* surges to the cable plant, while the voltages recorded in the field were truncated, with only values above 60 volts being recorded. The desired distribution is, therefore, the conditional probability that a voltage peak exceeds T volts given that it is greater than 60 volts. This may be accomplished by shifting the distribution function to the left until the probability of exceeding 60 volts is equal to one.

Thus,

$$P(V > T) = e^{-aT}e^{60a},$$

where $T \geqq 60$ volts. Therefore,

$$P(V > T) = e^{0.565}e^{-0.0094T} = 1.76e^{-0.0094T}.$$

This expression provides the probability distribution presented in Fig. 15.

To estimate the number of surges per thunderstorm day which exceed any given amplitude, it is necessary also to determine the average num-

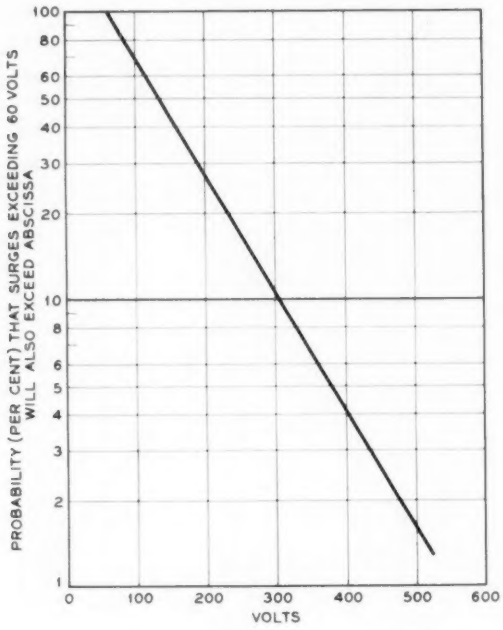


Fig. 15 — Probability distribution of all surges exceeding 60 volts in buried and aerial cable.

ber of surges induced in a cable per thunderstorm day. An estimated value was computed by counting the total number of recorded surges over the entire study period and dividing by the number of thunderstorm days occurring in the area. Very good correlation was established between observed thunderstorm days (on film) and those reported by local weather stations.

A total of 36 thunderstorm days was recorded in the vicinity of the buried cable at Griffin and six thunderstorm days were recorded in the vicinity of the aerial cable at Buford. These 42 thunderstorm days recorded for the two test locations accounted for a total of 1220 surges, for an average of 29 surges per thunderstorm day.

The actual number of surges induced in a cable, however, is approximately 2.5 times this value or 73 surges per storm. This results from the fact that multiple surges occur in approximately half of all lightning strokes, and that the average number of surges in a multiple discharge stroke is about four. The interval between surges is approximately $\frac{1}{10}$ second.[1,2] The sweep speed of the test oscillographs (full scale deflection

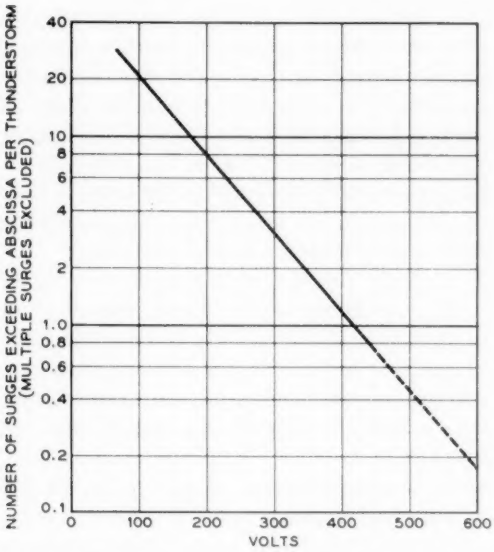


Fig. 16 — Surge distribution on buried and aerial cable, no protector blocks.

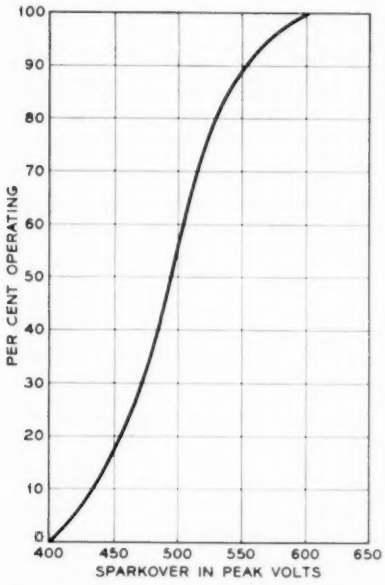


Fig. 17 — Sparkover distribution characteristics of 3-mil carbon block protectors.

of 1500 microseconds) was such that only the first surge was actually recorded in the field test. When evaluating apparatus that is vulnerable to repeated low-amplitude voltage surges, the occurrence of these multiple-stroke discharges must be considered. Lightning surge failure of most apparatus, however, is caused by the inability of the apparatus to handle the surge energy. Since subsequent surges in a multiple-stroke discharge generally have a lower energy content, apparatus vulnerable to surge energy will be most susceptible to failure on the initial surge, eliminating the need to test for the effects of multiple discharges.

Fig. 16 presents a plot of the expected distribution of surges (excluding multiple surges) on buried or aerial cable plant as a function of peak voltage in the absence of associated 3-mil protector blocks. The effect of protector blocks is to reduce the probability of observing voltages in the range of 400 to 600 volts. The distribution of protector block operation as a function of peak voltage had been determined, and is presented in Fig. 17. When this probability distribution is applied to the probable incidence of various surge voltages in cable plant the curve shown in Fig. 4 results. Fig. 4 gives the expected number of surges (excluding multiple surges) per thunderstorm exceeding any given voltage, up to the maximum sparkover voltage of 3-mil protector gaps.

REFERENCES

1. Wagner, C. F. and McCann, G. D., Induced Voltages on Transmission Lines, A.I.E.E. Trans., **61,** 1942, p. 916.
2. Sunde, E. D., Lightning Protection of Buried Toll Cable, B.S.T.J., **24,** 1955, p. 253.

A General Method of Applying Error Correction to Synchronous Digital Systems

By D. B. ARMSTRONG

(Manuscript received March 8, 1960)

A general method is presented for applying error correction to synchronous binary digital systems to improve reliability. It includes the familiar scheme of triplication and "vote taking" as a special case. In principle, the method permits the system to operate continuously, even when a fault is present or maintenance is being performed. An efficient maintenance routine, including rapid repair of faults, is an essential adjunct to the scheme if the potentially large increase in reliability made possible by error correction is to be realized.

The percentage redundancy needed to realize the scheme decreases as the complexity of the system to which it is applied increases, but may amount to triplication of equipment even for moderately large systems. The paper describes some error-correcting codes to implement the scheme, discusses error-correcting circuits in a general way, indicates how to estimate the redundancy, and presents a formula for determining the reliability improvement obtainable with a particular maintenance routine. In a companion paper,[1] D. K. Ray-Chaudhuri develops a general theory of minimally redundant codes for this application.

I. INTRODUCTION

This paper describes a general method of applying error correction to synchronous digital data systems. It includes, as a special case, the well-known scheme of triplication with vote taking.[2] Since the scheme employs error-correcting codes, it is capable of detecting errors as well as correcting them. Hence, maintenance personnel can be alerted as soon as a fault occurs. Also, it has the property of enabling the system to which it is applied to continue to function correctly even when faults are present and maintenance is being performed, provided all the faults are confined to any one of the several subunits which comprise the sys-

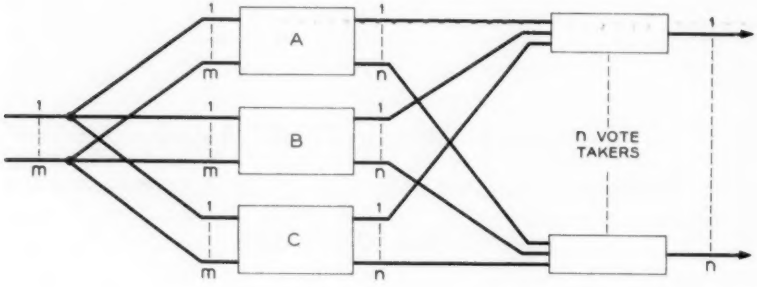


Fig. 1 — Error correction by triplication and vote taking.

tem. Therefore, if faults are found and repaired quickly, so that they do not accumulate to the point where the resulting errors are beyond the error-correcting capabilities of the scheme, the system may be kept in continuous operation for much longer periods than could the equivalent system without error detection and correction.

In this regard, it is estimated that the "mean life"* of the system with error correction can be made several thousand times as long as that of the equivalent nonredundant system, provided faults are repaired sufficiently soon after their occurrence. Such potentially vast increases in reliability depend of course on the availability of rapid diagnostic and fault-repair facilities. Conversely, in the absence of maintenance the mean life of the redundant system will in general be less than that of the nonredundant system. Hence the scheme is not usefully applicable to a system which must operate in an environment where rapid fault repair is impossible — in such situations some other method of building in reliability, such as microlevel redundancy,[3] would be necessary.

In comparison with triplication and vote taking, our procedure will permit more precise localization of faults. Also, for large systems it should result in less over-all equipment redundancy. For small systems, however, an equipment advantage may not always be realized. Since the triplication scheme is fairly well known, we shall start by describing it, but from a slightly different point of view, which shows how it appears as a special case of our procedure.

Fig. 1 shows a system, A, with m inputs and n outputs, and two exact replicas of the system, B and C. Corresponding output wires from A, B and C are fed to "majority" circuits, or vote takers, each of whose out-

* The mean life of a system is here defined as its mean time to failure, assuming it is in perfect condition at the start.

puts agrees with the majority, i.e., with any two or all three, of the inputs which are in agreement. Thus, the system corrects for errors which are confined to the outputs of any one of systems A, B or C.

We may consider the outputs from A to carry information bits, and those from B and C to carry check bits, which generate Hamming[4] single error-correcting codes, in the following manner. The matrix below displays the output bits from A, B and C in a matrix consisting of three rows, each row having n entries:

output #1	output #2	output #3		output #n	
↓	↓	↓		↓	
$A_1 \rightarrow$ O	O	O	···	O	← outputs from system A
$B_1 \rightarrow$ X	X	X	···	X	← outputs from system B
$C_1 \rightarrow$ X	X	X	···	X	← outputs from system C

O = information bit
X = check bit

Alternatively, the matrix may be thought of as displaying n columns, each with three entries consisting of one information bit and two check bits. For example, the first column contains the information bit A_1, and check bits B_1 and C_1. Two parity checks are constructed from this column; bits A_1 and B_1 satisfy the parity relation

$$A_1 \oplus B_1 = 0,$$

where $\oplus$ represents the sum modulo 2. Bits A_1 and C_1 satisfy the parity relation

$$A_1 \oplus C_1 = 0.$$

These relations merely state that, when the complete system is operating correctly, both B_1 and C_1 will have the same value as A_1.

This coding has the ability to detect any single error in column 1, and moreover tells us which bit is in error, so that corrections can be performed. Therefore, in particular, this scheme permits the correction of any pattern of errors which is confined to a single row of the matrix. Since faults which are confined to one of the systems A, B, or C can cause errors on the outputs of that system alone, this error-correcting scheme will permit the over-all system to operate correctly even when any one of the three systems comprising it is faulty, or is disabled for maintenance purposes.

Obviously, this particular coding is inefficient, because two check bits

are needed for each information bit. It is to be hoped that the use of more efficient codes will result in less equipment redundancy, primarily because fewer check bits will have to be generated. The problem then is to find a way of organizing a system so as to permit error correction with more efficient codes. A scheme for doing this will now be described.

II. A GENERAL SCHEME FOR ERROR CORRECTION

Suppose system A of Fig. 1 is designed so that it breaks down into a number, say r, of electrically independent subunits, each subunit carrying not more than p of the n system outputs, as shown in Fig. 2.

A fault or faults that is confined to any one subunit can at most cause errors on the outputs of that subunit. Therefore, consider the following matrix, in which the outputs of each subunit are displayed in a separate row, with p entries per row. There are $(q - r)$ additional subunits shown in Fig. 2; these provide k check bit outputs:

$$
\left.\begin{array}{l}
\left.\begin{array}{l}
\left.\begin{array}{llll} O & O & \cdots & O \leftarrow \text{outputs from subunit } 1 \\ \vdots & \vdots & & \vdots \\ O & O & \cdots & O \leftarrow \text{outputs from subunit } r \end{array}\right\} r \text{ rows} \\
\left.\begin{array}{llll} X & X & \cdots & X \leftarrow \text{outputs from subunit } r+1 \\ \vdots & \vdots & & \vdots \\ X & X & \cdots & X \leftarrow \text{outputs from subunit } q \end{array}\right\} (q-r) \text{ rows}
\end{array}\right\} q \text{ rows}
\end{array}\right.
$$

Since faults in a single subunit affect only a single row of the matrix, we may, for example, apply Hamming single error-correcting codes on a

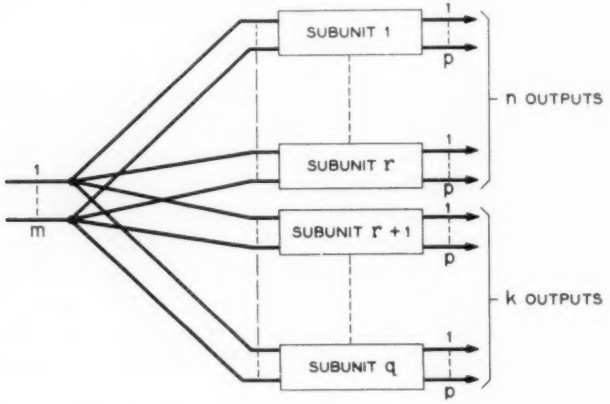


Fig. 2 — Breakdown of system A into subunits.

per-column basis. Thus, if $r = 4$, we need only three check bits per column to provide Hamming single-error correction, and the code redundancy is much less than in the triplication scheme ($\frac{3}{7}$ instead of $\frac{2}{3}$).

Actually, Hamming codes are not the most efficient that could be used for this error-correcting scheme, because they do more than is required. Specifically, they permit correction of any single error per column of the matrix, even though these errors may not be confined to a single row. If we apply the further restriction that all errors be confined to a row, then more efficient codes are possible and are described later.

We now wish to show how it is possible to break down a system into electrically independent subunits. Digital systems may be classified into two types: those which perform only combinational logic (have no memory), and those which perform sequential logic (have memory). The latter type is of more interest, but it is useful to deal with the former first. We assume throughout that the data on the input wires are not in error, and that faults in the system do not cause errors on the input wires.

Suppose then that the r subunits in Fig. 2, which produce the n system outputs, consist entirely of combinational logic. It is evident that the system can be broken down into such subunits because, for example, each output can be realized by designing a separate combinational logic circuit which generates the appropriate Boolean function of the m input variables. Alternatively, some savings in logic elements may be possible by designing multifunctional logic circuits, each generating only the p outputs of a single subunit.

To provide check outputs, additional subunits are needed, and are designated $(r + 1)$ through q in Fig. 2. To design these, it is necessary to be able to express each check output as a Boolean function of the m input variables. This can be done because the structure of the error-correcting code will specify each check output to be the sum modulo 2 of some set of information outputs, and since the latter are known functions of the inputs, we can therefore express the check outputs directly as functions of the inputs. We may, of course, work with truth tables instead of functional representations.

In the case of sequential logic, a complication is introduced which may be explained with the aid of Fig. 3. In this figure, a sequential system is represented as consisting of two major units. Unit 1 consists entirely of combinational logic and unit 2 consists entirely of memory.*

* Some authors replace the memory elements by unit delay elements. See for example, Fig. 1 of Unger.[5] His paper deals with asynchronous circuits, whereas we are treating synchronous circuits of the type designated "*PP*" by Cadden.[6]

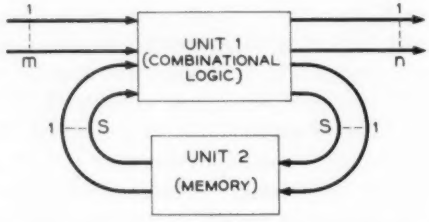


Fig. 3 — A possible configuration for a finite-state sequential system.

The combinational logic generates two sets of outputs:

(a) The n system outputs;

(b) s "feedback" outputs which provide the inputs to the memory unit.

The s outputs of the memory unit, in conjunction with the m system inputs, comprise the inputs to the combinational logic unit.

Suppose that unit 1 is designed as r electrically independent subunits. In general, the $(m + s)$ inputs to unit 1 will feed all r subunits. A fault in a single subunit will cause errors on the output of that subunit, and these will feed back via the memory unit to the inputs of some or all of the other subunits. Hence, in a few cycles of operation it is possible that the outputs of all subunits will be in error because of a fault in just one subunit. This situation can be remedied by applying error correction to some or all of the s feedback wires in addition to the n system output wires. These additional corrections should be made between the outputs of unit 2 and the inputs of unit 1, in order to correct errors caused by faults in unit 2 as well as in unit 1.

Alternatively, it is possible to design the system so as to avoid correcting the internal feedback wires, and yet insure that a fault affects not more than p of the n system outputs. For example, instead of breaking down the system into r subunits, one could replicate the system r times and utilize only outputs $1, 2, \cdots, p$, from the first replica, outputs $p + 1, p + 2, \cdots, 2p$, from the second replica $\cdots$ and outputs $n - p + 1, n - p + 2, \cdots, n$, from the rth replica.

No doubt this alternative realization could be achieved without using r complete replicas of the system. However, the necessary design procedures are not well formulated and the resulting equipment redundancy is difficult to estimate. In contrast, the design procedure for the first mentioned method is straightforward, its redundancy is easier to estimate and, at least with present devices and techniques, it appears to result in considerably less over-all redundancy. Therefore, in the remainder of the paper we shall assume that the first method is to be used.

Accordingly, the correction of a sequential system requires that unit 1 of Fig. 3 be designed as r independent subunits and that it be augmented by a combinational logic unit which generates k check outputs, where k is large enough to provide the necessary parity checks for correcting $(n + s)$ wires (we assume that all s feedback wires may require correction). As in the purely combinational case, each check bit can be expressed as a sum modulo 2 of an appropriate subset of the n outputs of unit 1 and the s outputs of unit 2, and since these are known Boolean functions of the $(m + s)$ inputs to unit 1, each check bit output can likewise be expressed as a Boolean function of these same inputs.

It is of course necessary that the check bit logic circuits also be designed as independent subunits with not more than p outputs per subunit.

III. ERROR-CORRECTING CODES

Before discussing specific codes, we wish to establish lower bounds on the number of check bits, k, needed to fulfill our error-correcting requirements. Specifically, referring to the matrix above for the outputs from r subunits of system A, we ask what minimum value of k is required to permit correction of every possible pattern of errors in any single row of the q rows.

Actually, two lower bounds are applicable. The first bound, which is also the larger of the two when $q > (2^p + 1)$, p being the number of entries per row, is easily derived as follows: Observe that the number of possible error patterns in a single row is $(2^p - 1)$, if we exclude the no-error pattern. Therefore, the total number of error patterns in all q rows is $q(2^p - 1)$. Obviously, k must be large enough to permit as many "parity failure" patterns as there are error patterns. This requires that k satisfy the inequality

$$(2^k - 1) \geqq q(2^p - 1).$$

That is,

$$k \geqq [\log_2 (q2^p - q + 1)], \tag{1}$$

where the square bracket denotes the smallest integer which is

$$\geqq \log_2 (q2^p - q + 1).$$

The second lower bound, which is larger than the first when $q \leqq (2^p + 1)$, and which is therefore of greater practical significance, is:

$$k \geqq 2p. \tag{2}$$

It is derived by determining the maximum number of code words that

can be chosen out of a set of 2^{pq} binary words of lengths pq, and which fulfill the specified error-correcting requirements. Its derivation is relegated to the Appendix.

Surprisingly, it was found not too difficult to construct codes for most values of p and q in the range $2 < p < 10$, $3 < q < 9$ which achieved the appropriate lower bound.

Subsequent to the work described here, Ray-Chaudhuri[1] developed a general theory of minimally redundant codes for this application. However, it will not be out of place to exhibit here some of the codes previously derived, since they are also minimally redundant, and since the error-correcting equipment required to implement either them or the Ray-Chaudhuri codes is of the same general character and complexity. Three families of codes* are exhibited below in matrix form, corresponding to three values of p, as follows:

Family 1: $p = 2$; $q = 3$, 4 and 5.
Family 2: $p = 3$; $q = 5$, 6, 7, 8 and 9.
Family 3: $p = 4$; $q = 4$, 5, 6 and 7.

Family 1 ($p = 2$)

1	2
X	X
3	4
X	X
14	23
O	O
13	124
O	O
24	123
O	O

Family 2 ($p = 3$)

1	56	24
X	O	O
3	15	25
X	O	O
5	146	36
X	O	O
6	34	2
X	O	X
4	126	35
X	O	O
13	46	1245
O	O	O
16	456	23
O	O	O
14	125	236
O	O	O
45	136	256
O	O	O

Family 3 ($p = 4$)

1	5	23	46
X	X	O	O
2	6	38	57
X	X	O	O
3	7	14	58
X	X	O	O
4	8	12	67
X	X	O	O
18	25	36	47
O	O	O	O
16	24	58	378
O	O	O	O
13	27	56	478
O	O	O	O

* These codes were constructed by George Allen at Bell Telephone Laboratories in Summer, 1959.

The set of digits beside each bit position indicates the parity groups which that bit enters. For example, in family 1, the digit set attached to the last bit in the last row is 123, indicating that this bit enters parity check groups 1, 2 and 3. As a further illustration, the bits in family 1 that are labeled 1, 14, 13, 124 and 123 enter parity group 1; their sum modulo 2 is zero when there are no errors. The bit positions which carry only a single digit are check bit positions. They are denoted by X's and the information bits are denoted by O's. The matrix displayed for family 1 is a 2×5 matrix; however, if a 2×4 or a 2×3 matrix is desired, one omits respectively the last row or the last two rows. Similar remarks apply to families 2 and 3.

Several remarks should be made at this point. First, observe that in the original matrix we represented the check bits as being located entirely in the last $(q - r)$ rows. However, this is not necessary; the check bits may appear along with information bits in some or all rows, as is the case in the three matrices above. The arrangement is dictated by the structure of the code, but bits may be permuted within each row with impunity.

Secondly, it is possible to delete information bits from any row without destroying the utility of a code; in such cases, the deleted bits will be omitted from the parity checks in which they would normally participate.

Finally, we observe that the three matrices above provide only for values of $p \leqq 4$ and $q \leqq 9$. If for any reason we wish to form matrices with $p > 4$, or $q > 9$, this may be done by building up the over-all matrix, either vertically or horizontally, or both, from several of the above matrices. Thus a wide variety of equipment arrangements can be accommodated. However, if we build up vertically, we will sacrifice minimal code redundancy. For example, if we form a matrix with 3 columns and 18 rows by using two matrices of family 2, we shall have included $2 \times 6 = 12$ check bits, whereas the minimum is given by bound 1, namely

$$[\log_2 (18 \times 2^3 - 18 + 1)] = 7 \text{ bits.}$$

IV. ERROR-CORRECTING CIRCUITS

A discussion of error-correcting circuits is included here to indicate roughly the amount of equipment involved in error correction, and to provide a basis for a maintenance routine which is proposed later.

It was explained in Section II that, with the present scheme, it is necessary to apply error correction either to the n system outputs of a purely combinational system, or to the $(n + s)$ outputs and feedback

connections of a sequential system. To do this, k parity check bits must be generated, where k is determined by the structure of the error-correcting code employed. Therefore, the inputs to the error-correcting circuits consist of $(N + k)$ wires, where $N = n + s$ ($s = 0$ for a purely combinational system) and k wires carry the check bits. The output of the error-correcting circuits consists of N wires which carry the corrected versions of the corresponding N inputs.

We shall want to distinguish between the correcting circuits and the circuits which are corrected or correctable. The latter comprise the set of q subunits which generate the $(N + k)$ outputs; for example the circuits in Fig. 2. Therefore, we shall hereafter refer to the q subunits as "the system," without modifier.

The error correcting-circuits may be considered to perform the following three functions:

1. Reconstruct the parity checks to determine which, if any, have failed.

2. From the pattern of parity check failures, determine which of the $(N + k)$ input wires are carrying erroneous bits.

3. Correct that subset of the N wires which are carrying erroneous bits. Erroneous check bits do not require correction.

Circuits to perform the above tasks may be realized in several ways, and with varying degrees of redundancy to assure reliability. At one extreme, the error-correcting circuits could be nonredundant, in which case an efficient preventive maintenance routine would be required to insure that they perform for long periods without error. Alternatively they could be built with microlevel redundancy, in which case preventive maintenance would again be necessary but would be applied less frequently. A third alternative would be to make some or all of the error-correcting circuits "self-error-detecting." Those parts which were self-error-detecting would be subjected to maintenance only when a fault was detected; the parts which were not self-error-detecting would require preventive maintenance.

As a fourth alternative, it might be attempted to make the error-correcting circuits completely self-error-correcting. However, a simple heuristic argument can be given which indicates that it is impossible to achieve this goal.

Fig. 4 shows a block diagram of a proposed error-correcting circuit, designed according to the third alternative above. Box 1 in Fig. 4 contains the units which perform functions 1 and 2 above. Box 2 performs function 3 above, and also an error-sensing and alarm function. For simplicity, sets of wires in this figure are represented by single directed

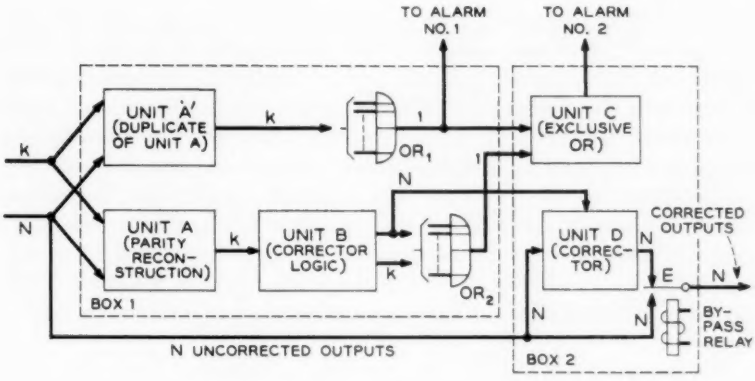


Fig. 4 — A proposed error-correcting circuit.

lines with an associated symbol indicating the number of wires. Box 1 is designed to be self-error-detecting, but box 2 is not. A detailed description of the operation of the circuits in Fig. 4 does not contribute significantly to an understanding of the over-all scheme, and so is omitted.

V. RELIABILITY ANALYSIS

We shall first describe a maintenance routine which is appropriate to the particular mode of error correction realized by the circuits of Fig. 4. We shall then apply an approximate formula which indicates the improvement in reliability of the redundant error-correcting system over the original nonredundant system when our particular maintenance routine is employed. For brevity, derivation of this formula is omitted.

The maintenance routine is as follows. If a fault occurs in some subunit of the system, unit D under control of box 1 corrects the resulting errors. Simultaneously, box 1 operates alarm 1. The faulty subunit is then located and repaired as quickly as possible. In principle, the system can continue to operate correctly even when the faulty subunit is being replaced or repaired, provided a fault does not develop in another subunit or in the error-correcting circuits during the repair of the original fault.

If box 1 fails, alarm 2 is operated, and possibly alarm 1 also, and simultaneously relay E is switched to the bypass position, thus causing essentially no interruption in system operation. Box 1 must also be repaired quickly, because if a fault occurs in the system during the repair of box 1, the system will fail, since it is not now being error cor-

rected. If units C or D fail, they do not operate an alarm since they are not provided with error detection. The probability of failure of these units must, therefore, be minimized by a preventive maintenance routine. To minimize the outage times of units C and D, two copies of each could be provided, one operating and one standby. They would be interchanged at regular intervals T, and preventive maintenance would be applied to the units currently in the standby condition. In this way, the system would not lose its error correcting capability during preventive maintenance. Relay E must be switched to the bypass position to permit continuous operation while unit C or D is being replaced. We assume that the relay switching time is short enough so as not to interrupt system operation.

We now wish to determine a quantitative measure of the reliability improvement of the maintained error-correcting system over that of the nonredundant system. A useful measure is the ratio of their respective mean lives; namely, $R_L = L_r/L_i$, where L_r is the mean life of the redundant system and L_i is the mean life of the nonredundant system. In general, the derivation of R_L is quite complicated, but by making suitable simplifying assumptions we can obtain an approximate formula which is useful. These assumptions are:

1. All components have an exponential survival probability function and the same mean life, which is taken to be the time unit. Therefore, the survival probability of any component is exponential $(-t)$.
2. Components fail independently.
3. Failure of any component in the nonredundant system causes that system to fail.
4. The bypass relay has zero probability of failure.
5. The times taken to repair faulty circuits are assumed constant.

These and other parameters are now defined:

Δ_1 = repair time of a subunit of the system,

Δ_2 = repair time of box 1,

Δ_3 = time that unit C or unit D is removed from circuit when being exchanged with its standby,

T = time interval between successive replacements of unit C or D by its standby

n_1 = number of components in each subunit of the system,

n_2 = number of components in box 1,

n_3 = number of components in box 2,

q = total number of subunits in the system,

r = number of subunits which provide the N system outputs.

From these assumptions plus some others concerning the relative

values of the above parameters, we can obtain the following formulas:

$$R_L = \frac{n_1 q}{2\delta}, \tag{3}$$

where

$$\delta = \tfrac{3}{2} n_3 T + r n_1 \frac{\Delta_3}{T} + q n_1 [(q-1) n_1 \Delta_1 + n_2(\Delta_1 + \Delta_2)]. \tag{4}$$

As an application of (3) and (4), consider a system having parameter values as follows:

time unit = mean component life = 1000 years,

$\Delta_1 = \Delta_2$ = one-half hour = $\frac{1}{2} \times 10^{-7}$ units (time to repair a system subunit or box 1),

Δ_3 = 8 seconds = 2×10^{-10} units (time to replace unit C or unit D by its standby),

T = one month = 10^{-4} units (maintenance interval for units C and D),

$n_1 = n_3 = 333$ components,

$n_2 = 666$ components,

$q = 6$,

$r = 4$.

Substituting these values in (3) and (4) results in R_L equal to 2900. That is, the mean life of the redundant system is 2900 times that of the nonredundant system. Actually, this figure could be improved if we reduced the repair times Δ_1 and Δ_2 from one-half hour to, say, five minutes. Such a reduction would be possible if the system were built of small modular packages and a highly automated diagnostic routine were available to locate faults in the order of a minute or less. Thus, a fault could be pinpointed to one or two particular packages, and these packages could be replaced immediately by good standby packages, thus permitting correction of the fault in minutes. The faulty packages could then be tested and repaired in more leisurely fashion, and this latter time would not be chargeable to Δ_1 or Δ_2.

Therefore, it appears that, with an efficient maintenance routine, the mean life of the error-correcting system can be several thousand times that of the nonredundant system.

VI. REDUNDANCY AND SPEED PENALTIES

It would be desirable to estimate the amount of equipment redundancy in the scheme described, and any attendant reduction in the operating speed of the system. The equipment redundancy cannot be specified

in simple terms, but the contributing sources can be delineated and roughly evaluated. They are:

(a) the circuits which generate the check bits;

(b) the circuit redundancy which results from designing the system as q independent subunits;

(c) the error-correcting circuits.

The redundancy contribution of item (a) can reasonably be estimated to be in the ratio k/N to the amount of equipment in the original non-redundant system. That is, if the amount of equipment in the non-redundant system is treated as one "unit" of equipment, the amount of equipment needed to generate the check bits would be roughly k/N "units." As indicated by the codes in Section III, values of k/N as small as $\frac{1}{3}$ are achievable for $N \geq 18$. The assumption underlying this estimate is that the amount of circuitry required to generate each check output is the same as the amount required to produce each original system output.

The redundancy contribution of item (b) is believed to be insignificant compared to the other contributions, especially for large systems, provided optimal design techniques are employed. The contribution of item (c) is by far the largest, and is the most difficult one to estimate. It depends on the type of logic technology employed, on the amount of time delay that the error-correcting function is permitted to introduce, and to some extent on the particular error-correcting code used. The following estimates may suggest an order of magnitude for item (c), in the particular case of the correcting circuits proposed in Fig. 4, and assuming the use of diode logic. Based on "paper" designs of these circuits, the author estimated that the correcting circuits might require roughly 60 to 70 "equivalent" diodes per wire corrected (the number of wires corrected is N). Transistors were counted as equivalent to two diodes, resistors etc. were not counted.

This estimate assumes that units A and A′ of Fig. 4 both employ $(c - 1)$ EXCLUSIVE OR circuits per parity check over c bits, and that unit B of Fig. 4 is realized with two-stage logic. Thus, if these error-correcting circuits were to be applied to a system which, in its non-redundant form, was realizable with 30 to 35 "equivalent" diodes per wire corrected, it is evident that the correcting circuits would comprise two "units" of equipment. This ratio is not considered to be unrealistically high.[1]

In general, therefore, it is to be expected that the scheme in question may introduce an amount of redundancy equivalent to at least triplication of the original equipment. However, it has the potential of being

less redundant than the triplication and vote-taking scheme, in which generation of the check bits alone results in triplication and the correcting circuits are additional.

In this connection, it is remarked that the triplication scheme is usually thought of as including relatively simple vote-taking circuits which perform corrections but are incapable of detecting errors and operating appropriate alarms. These latter features would have to be included in order to render the scheme truly comparable to the more general error-correcting procedure described here.

In principle, it appears that the redundancy penalty might be made to decrease monotonically as the systems to which error correction is applied becomes increasingly complex, provided the following two assumptions are valid:

(a) as the number N of corrections is increased, the coding efficiency also increases; that is, k/N becomes smaller;

(b) the amount of equipment per correction in the original system increases faster than the amount of equipment per correction in the correcting circuits.

Assumption (a) is realizable, but (b) cannot be verified. Indeed, (b) may be plausible only provided the correcting circuits use an increasing number of logic stages, which can be expected to result in an increase in time taken to perform corrections; that is, an increasingly severe speed penalty is imposed.

In this regard, the parity check circuits referred to earlier in this section require $2 \times [\log_2 c]$ logical stages. (The square bracket denotes the smallest integer which is equal to or greater than $\log_2 c$.) For the special codes described in Section III, the number of bits per parity check, c, is typically equal to q, the number of subunits in a system, and q must increase in order to increase the coding efficiency. It therefore follows that greater coding efficiency can be achieved only at the expense of greater delay in the corrector, or more complex correcting circuits, or both, and a compromise must be reached.

Finally, a remark should be made concerning the impact of this scheme on the over-all design of a sequential system when the method which requires both feedback and output corrections is used. To minimize the number of corrections necessary, the number of feedback and output wires should be kept to a minimum. The designer usually is able to exercise some control over both. In particular, there are roughly as many feedback connections in a sequential system as there are binary memory elements; therefore it would be desirable to minimize the number of memory elements. At present, large systems are frequently

designed with many more memory elements than necessary, presumably because this results in simpler design procedures. It may, therefore, be desirable to find methods which lead to designs having nearly minimal numbers of memory elements, in order to make our error-correcting procedure more attractive.

VII. SUMMARY

We have described an error-correcting scheme which is generally applicable to synchronous digital systems, and which includes the triplication and vote-taking scheme as a special case. It permits systems to which it is applied to operate continuously even when faults are present and maintenance is being performed. The scheme can lead to very large increases in system reliability, but only if augmented by a maintenance routine which effects rapid repair of faults.

Two types of error-correcting codes have been discussed, Hamming codes and special codes. The Hamming codes are universally applicable, but are not minimally redundant in this application. The special codes are minimally redundant but not universally applicable, in that they have not been developed for a large range of values of p and q.

The equipment redundancy required to implement the scheme may be equivalent to at least triplication for moderately large systems, but should be less for more complex systems. It is not specifiable in simple terms and can be determined accurately only by carrying through the detailed design of the specific systems. Such detailed applications have not yet been made.

APPENDIX

Proof That 2p Is a Lower Bound on the Number of Check Bits

We shall derive this bound by showing that an upper bound on the number of code words of length pq which satisfy our error-correcting criterion is $2^{(q-2)p}$ words. This implies that the maximum number of bits which can be assigned values arbitrarily is $(q - 2)p$ bits. The remaining bits must be check bits; therefore, a lower bound on the number of check bits is $qp - (q - 2)p = 2p$ bits.

Proof That an Upper Bound on Number of Code Words Is $2^{(q-2)p}$

It is useful to think of the pq bits which comprise a code word as being arranged in a single row, with each successive block of p bits being replaced by a single symbol, D_i, which can take on any one of the 2^p

different values. In this alternative representation, a typical q-symbol word would be

$$D_1D_2D_3 \cdots D_q .$$

In terms of this representation, our error-correcting criterion requires that an error in any one of the q symbols be correctable. This implies that admissible code words must differ in more than two symbol positions. For, consider the following two words which differ only in the first two symbol positions:

$$\text{word 1:} \quad D_1D_2D_3 \cdots D_q ,$$

$$\text{word 2:} \quad D_1'D_2'D_3 \cdots D_q .$$

It is possible for an error in the first symbol of word 1 and in the second symbol of word 2 to cause both to become, for example, the word

$$D_1'D_2D_3 \cdots D_q .$$

Hence, we cannot determine whether word 1 or 2 was the correct word; therefore, admissible code words must differ in more than two symbol positions.

Let S be the set of all q-symbol words. There are 2^{qp} such words. Partition S into disjoint subsets S_i, $i = 1, 2, 3, \cdots$, such that two elements of S belong to the same subset if they are identical in the last $(q - 2)$ symbol positions. Thus there are as many subsets as there are truncated words $D_3D_4 \cdots D_q$, namely, $2^{(q-2)p}$ subsets, and each subset contains 2^{2p} elements.

Now arbitrarily choose a (q-symbol) word from subset S_1 to be a code word. Then no other words from subset S_1 can be chosen as code words, because any two words from S_1 differ only in the first two-symbol positions. By the same argument, at most one word can be selected as a code word from S_2, etc. Therefore, there cannot be more code words than subsets. Hence, an upper bound on the number of code words is $2^{(q-2)p}$.

REFERENCES

1. Ray-Chaudhuri, D. K., this issue, p. 595.
2. Von Neumann, J., Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components, in Shannon, C. E. and McCarthy, J., eds., *Automata Studies*, Annals of Math. Studies, No. 34, Princeton Univ. Press, Princeton, N. J., 1956, pp. 44–98 (particularly Section 8.3).
3. Moore, E. F. and Shannon, C. E., Reliable Circuits Using Less Reliable Relays, J. Frank. Inst., **262**, 1956, pp. 191; 281.
4. Hamming, R. W., Error Detecting and Error Correcting Codes, B.S.T.J., **29**, 1950, p. 147.
5. Unger, H., Hazards and Delays in Asynchronous Sequential Switching Circuits, I.R.E. Trans., **CT-6**, 1959, p. 12.
6. Cadden, W. J., Equivalent Sequential Circuits, I.R.E. Trans., **CT-6**, 1959, p. 30.

On the Construction of Minimally Redundant Reliable System Designs

By D. K. RAY-CHAUDHURI

(Manuscript received September 7, 1960)

Several authors have considered the possibility of increasing the reliability of large and complex binary digital systems by introducing some redundancy in the system. In a companion paper, Armstrong[1] proposes a scheme for applying error correction to a synchronous digital system. In this paper we develop a general mathematical theory for generating minimally redundant error-correcting codes for the scheme in question. This results in what are called "minimally redundant reliable systems." The problem of constructing minimally redundant reliable systems whose output is free of error when there is a fault in at most one block of the system is completely solved. An example is considered in detail showing how the mathematical theory can be actually applied.

I. INTRODUCTION

In complex binary digital systems employing a large number of blocks of electrical equipment it often is difficult to ensure a sufficient level of reliability of each single block of equipment. An attempt to attain the desired degree of reliability by improving the reliability of each block may prove to be uneconomical. On the other hand, by introducing some redundancy in the system, it is possible to construct highly reliable complex systems, even though each single block is not as highly reliable. Moore and Shannon,[2] Tryon,[3] Von Neumann,[4] Lofgren[5] and Armstrong[1] have considered the problem of constructing reliable system designs. In this paper a general mathematical theory has been developed for the construction of minimally redundant reliable system designs, based on the scheme outlined by Armstrong.[1] This theory is closely related to the theory of error-correcting codes. The problem of constructing minimally redundant system designs whose outputs will be free of error whenever there is fault in at most one block of the system is completely solved in this paper.

II. FORMULATION OF THE PROBLEM

Suppose there are m binary input variables $X_1, X_2, \cdots, X_m$. Let B_m denote the set of 2^m m-place binary sequences. Every set of values of the m binary input variables will be regarded as an element of B_m. Any mapping of B_m into B_1 will be called a *Boolean function* of the m input variables $X_1, X_2, \cdots, X_m$. For the sake of brevity, the collection of m input variables will be denoted by X. Let

$$f_{11}, f_{12}, \cdots, f_{1p}, \qquad f_{21}, f_{22}, \cdots, f_{2p}, \qquad f_{k1}, f_{k2}, \cdots, f_{kp}$$

be pk Boolean functions of the m binary variables $X_1, X_2, \cdots, X_m$. Our problem is to construct a system which will synthesize the pk Boolean functions with a high degree of reliability. The system uses blocks of electrical equipments each of which can synthesize p Boolean functions. For the sake of brevity, a collection of p Boolean functions, will be called a *Boolean p-function*. Thus $f_i = (f_{i1}, f_{i2}, \cdots, f_{ip})$ is a Boolean p-function. Any Boolean p-function is a mapping of B_m into B_p. Each block of our system synthesizes a Boolean p-function. Fig. 1 is a schematic diagram for the original nonredundant system.

The blocks act as units in the system. If there is a fault in a block, then some or all the p outputs of the block are erroneous. In other words, in the case of a fault a block will synthesize the corresponding Boolean p-function wrongly. Let V_p^1 denote the set of 2^p binary p-tuples. Then any Boolean p-function takes values on V_p^1. Let V_p^k denote the set of k-vectors $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_k)$, where each α_i is an element of V_p^1, $i = 1, 2 \cdots, k$. Let $f = (f_1, f_2, \cdots, f_k)$. Then f can be regarded as a mapping of B_m onto V_p^k. We shall define the addition of p-tuples as the usual mod 2 addition. For example, if $p = 3$, $\alpha_1 = (001)$ and $\alpha_2 = (101)$, then $\alpha_1 + \alpha_2 = (100)$. Let α and α' be two elements of V_p^k given by $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_k)$ and $\alpha' = (\alpha_1', \alpha_2', \cdots, \alpha_k')$. The sum $\alpha + \alpha'$

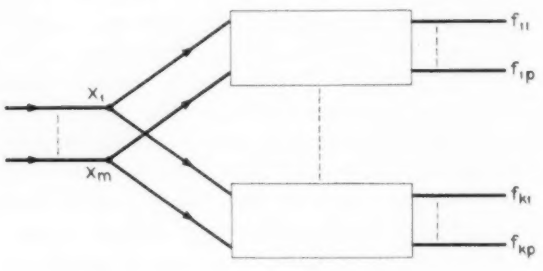


Fig. 1 — Original nonredundant system.

is defined to be the element $(\alpha_1 + \alpha_1', \cdots, \alpha_k + \alpha_k')$. The p-tuple $(00 \cdots 0)$ will be called the *null element* of V_p^1. The weight $\omega(\alpha)$ of the k-vector α is defined to be the number of nonnull elements among $\alpha_1, \alpha_2, \cdots, \alpha_k$. For any particular value X' of the input variables $f(X')$ is a vector in V_p^k. Suppose there are faults in t $(t < k)$ blocks. Then t of the functions $f_1, f_2, \cdots, f_k$ will be synthesized wrongly. Hence the output will be the vector $f(X') + \epsilon$, where $\epsilon = (\epsilon_1, \epsilon_2, \cdots, \epsilon_k)$ is a vector in V_p^k with weight t. While designing a system to synthesize the Boolean function f, one might require that whenever the number of faulty blocks is t or less, the output is error-free. One can achieve this by introducing some redundancy in the system, i.e., by synthesizing $(k + r)$ Boolean p-functions and adding a logical corrector unit to the system.

Suppose $\varphi_1, \varphi_2, \cdots, \varphi_n$ are n Boolean p-functions and C is a mapping of V_p^n onto V_p^k. We shall consider $\varphi = (\varphi_1, \varphi_2, \cdots, \varphi_n)$ as a function from B_m to V_p^n. For every value X' of X, $\varphi(X)$ is an element of V_p^n. Suppose the functions φ and C possess the property P stated below:

For every vector ϵ belonging to V_p^n with $\omega(\epsilon)$ not exceeding t and every value X' of the input variable X,

$$C(\varphi(X') + \epsilon) = f(X'). \tag{1}$$

The functions φ and C enable us to construct a system which will synthesize the k Boolean p-functions $f_1, f_2, \cdots, f_k$ free of error whenever the number of faulty blocks in the system is t or less. The n Boolean p-functions $\varphi_1, \varphi_2, \cdots, \varphi_n$ can be considered as a collection of np Boolean functions of m input variables. Therefore we can easily obtain the logical design of a system which will synthesize these np Boolean functions. This system will be called the *encoder subsystem*. Similarly, the function C can be considered as a collection of pk Boolean functions of np binary input variables, and therefore we can obtain a system which will synthesize these pk functions. This system will be called the *corrector subsystem*. The np outputs of the encoder subsystem will be the inputs of the corrector subsystem. Now it is easily seen that, because of the property P of the functions φ and C, whenever the number of faulty blocks in the encoder subsystem is t or less and the corrector unit is free of error, the pk outputs of the corrector subsystem will be

$$\begin{aligned} f(X) &= \{f_1(X), f_2(X), \cdots, f_k(X)\} \\ &= \{f_{11}(X), f_{12}(X), \cdots, f_{1p}(X), f_{21}(X), f_{22}(X), \cdots, f_{2p}(X), \cdots, \\ &\qquad f_{k1}(X), f_{k2}(X), \cdots, f_{kp}(X)\}. \end{aligned} \tag{2}$$

A schematic diagram for the whole system is given in Fig. 2. In view of

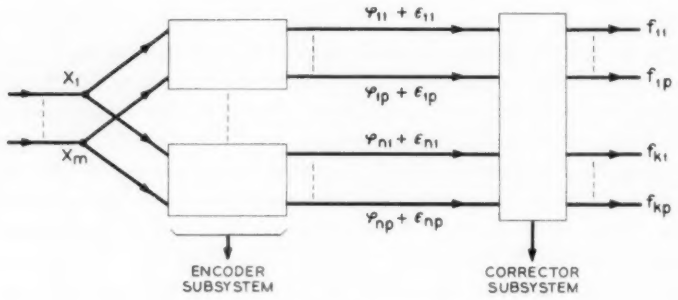


Fig. 2 — Whole system.

the above discussion, the following two definitions given below are meaningful.

Definition 1: The functions $\varphi_1, \varphi_2, \cdots, \varphi_n$ and C possessing the property P stated in (1) will be called a *reliable system design* of order t and redundancy $r = n - k$ for the k Boolean p-functions $f_1, f_2, \cdots, f_k$.

Definition 2: A reliable system design of order t and redundancy r_0 for the k Boolean p-functions $f_1, f_2, \cdots, f_k$ will be called *minimally redundant* if the redundancy r of any other reliable system design of order t for the same functions is not less than r_0.

In the present paper we have given a method of obtaining a minimally redundant system design of order 1 for any set of k Boolean p-functions for arbitrary k and p. System designs of higher order will be given in a subsequent paper.

We have used the redundancy r as a measure of the extra amount of equipment which has to be used for making the system reliable. And hence we seek the system which has minimum possible value of the redundancy r. It should be pointed out that we assumed that the corrector subsystem does not make any error at all. Therefore, to make the whole development practically feasible, it is imperative that either the amount of equipment necessary for the corrector subsystem is small in comparison to the amount of equipment necessary for the whole system, or that other steps be taken, such as are suggested in Ref. 1, to ensure reliability of the corrector system. We have not used any mathematical criterion to incorporate this requirement in the development of the theory.

III. LOWER BOUNDS ON THE REDUNDANCY r OF A RELIABLE SYSTEM DESIGN OF ORDER t

Consider two vectors α and α' belonging to V_p^n. The distance $d(\alpha,\alpha')$ between these two elements of V_p^n is defined to be $\omega(\alpha + \alpha')$. Thus if

$\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_n)$ and $\alpha' = (\alpha_1', \alpha_2', \cdots, \alpha_n')$, then $d(\alpha,\alpha')$ is equal to the number of integers i for which $\alpha_1 \neq \alpha_i'$, $i = 1,2,\cdots,n$. For example, if $p = 2$, $n = 3$, and $\alpha = (01,11,10)$ and $\alpha' = (01,10,00)$, then $\alpha + \alpha' = (00,01,10)$ and $d(\alpha,\alpha') = 2$. It can be easily checked that the distance defined above satisfies the three conditions of a distance function. We have seen that Boolean p-functions as defined in Section II can be considered as mappings of B_m into V_p^1. A Boolean p-function f_1 will be called a *nondegenerate* Boolean p-function if for any element α_1 of V_p^1, there is a value X' of the input variable X for which $f_1(X') = \alpha_1$. We shall assume that all Boolean p-functions appearing in our discussion are nondegenerate. In the following we have $s = 2^p$ and $n = k + r$.

Theorem 1: A necessary and sufficient condition that there exists a reliable system design of order t and redundancy r for the k Boolean p-functions $f_1, f_2, \cdots, f_k$ is that there exists a subset A of V_p^n containing s^k elements such that $d(\alpha,\alpha') \geqq 2t + 1$; $\alpha,\alpha' \in A'$, $\alpha \neq \alpha'$.

Proof: Necessity. Suppose there exists a reliable system design of order t. Let the encoder functions be $\varphi = (\varphi_1,\varphi_2,\cdots,\varphi_n)$ and the corrector function be C. For every value X' of the input variable X, $\varphi(X')$ is a vector of V_p^n. Consider the set

$$A = \{\varphi(X') \mid X' \in B_m\}.$$

Using the fact that the Boolean p-functions $f_1, f_2, \cdots, f_k$ are nondegenerate functions, it follows easily that the set A contains at least s^k vectors of V_p^n. Consider two distinct vectors α and α' of the set A. If possible, suppose $d(\alpha,\alpha') \leqq 2t$. Since $d(\alpha,\alpha') \leqq 2t$, we can find a vector ϵ of V_p^n such that $\alpha + \epsilon = \alpha' + \epsilon$ and $\omega(\epsilon) \leqq t$. Since $\omega(\epsilon) \leqq t$, we have

$$C(\alpha + \epsilon) = C(\alpha' + \epsilon) = \alpha = \alpha'. \tag{3}$$

Equation (3) contradicts the assumption that α and α' are distinct vectors of A. This completes the proof of necessity.

Sufficiency. Suppose A is a subset of V_p^n containing s^k elements and having the property that $d(\alpha,\alpha') \geqq 2t + 1$; $\alpha,\alpha' \in A$, $\alpha \neq \alpha'$. We set up a one-to-one correspondence between the s^k vectors of V_p^k and the s^k vectors of A. For every value X' of the input variables X, $f(X') = [f_1(X'),f_2(X'),\cdots,f_k(X')]$ is a vector of V_p^k and there is a corresponding vector α of V_p^n belonging to A. The encoder function $\varphi = (\varphi_1, \varphi_2, \cdots, \varphi_n)$ is defined by

$$\begin{aligned} \varphi(X') &= [\varphi_1(X'),\varphi_2(X'),\cdots,\varphi_n(X')] \\ &= (\alpha_1, \alpha_2, \cdots, \alpha_n) \\ &= \alpha, \end{aligned} \tag{4}$$

where α is the vector of V_p^n belonging to A and corresponding to the vector $f(X')$ of V_p^k. The corrector function C is defined in the following manner. Let $\gamma = (\gamma_1, \gamma_2, \cdots, \gamma_n)$ be an arbitrary vector of V_p^n. First we choose a vector α belonging to A such that $d(\gamma,\alpha) \leqq d(\gamma,\alpha')$, $\alpha,\alpha' \in A$. Let $\beta = (\beta_1, \beta_2, \cdots, \beta_k)$ be the vector of V_p^k which corresponds to α. Then we define

$$C(\gamma) = \beta. \tag{5}$$

Thus C is a mapping of V_p^n onto V_p^k. It is easy to check that the encoder function φ and the corrector function C defined above possess the property P stated in Section II. This completes the proof of sufficiency.

Theorem 2: If there exists a reliable system design of order t and redundancy r for k Boolean p-functions, then

$$s^r \geqq 1 + \binom{n}{1}(s-1) + \binom{n}{2}(s-1)^2 + \cdots + \binom{n}{t}(s-1)^t, \tag{6}$$

where $n = k + r$ and $s = 2^p$.

Proof: From Theorem 1, it is necessary that there exists a subset A of V_p^n with the property that

$$d(\alpha,\alpha') \geqq 2t + 1; \qquad \alpha,\alpha' \in A, \quad \alpha \neq \alpha'. \tag{7}$$

Let S_α denote the set of vectors γ of V_p^n with the property that $d(\gamma,\alpha) \leqq t$. It follows easily from (7) that, for any two distinct vectors α and α' of A, the sets S_α and $S_{\alpha'}$ do not have any common element. Let $S_\alpha^{(k)}$ denote the set of elements of V_p^n which have distance k from α, $k = 0,1,2,\cdots,t$. Obviously S_α is the union of the $(t+1)$ sets $S_\alpha^{(k)}$; $k = 0,1,\cdots,t$. $S_\alpha^{(k)}$ contains

$$\binom{n}{k}(s-1)^k$$

elements. Hence S_α contains

$$1 + \binom{n}{1}(s-1) + \binom{n}{2}(s-1)^2 + \cdots + \binom{n}{t}(s-1)^t$$

elements. There are s^k such nonoverlapping sets and the total number of elements of V_p^n is s^n. Hence we have

$$s^n \geqq s^k\left[1 + \binom{n}{1}(s-1) + \binom{n}{2}(s-1)^2 + \cdots + \binom{n}{t}(s-1)^t\right]. \tag{8}$$

Theorem 2 follows from (8).

Theorem 2 gives a lower bound on the redundancy r of a reliable

system design of order t for k Boolean p-functions. Theorem 2 is actually a generalization of a result of Hamming.[6]

Let $n_t(r)$ denote the maximum integer n for which there exists a reliable system design of order t and redundancy r for $k = n - r$ Boolean p-functions. For $t = 1$, the inequality (6) becomes

$$s^r \leqq \binom{n}{1} (s - 1).$$

Hence we have

$$n_1(r) \leqq \frac{s^r - 1}{s - 1}.$$

In Section V we shall show that

$$n_1(r) = \frac{s^r - 1}{s - 1}.$$

If there exists a reliable system design of order t and redundancy r for k Boolean p-functions, then $n_t(r) \geqq k + r$.

Lemma 1:

$$n_t(r + 1) \geqq n_t(r) + 1.$$

Proof: Suppose $n_t(r) = n$. Then there exists a reliable system design of order t and redundancy r for $k = n - r$ Boolean p-functions. Hence by Theorem 1 there exists a subset A of ${V_p}^n$ containing s^k elements with the property that $d(\alpha,\alpha') \geqq 2t + 1$; $\alpha,\alpha' \in A$, $\alpha \neq \alpha'$. To every vector $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_n)$, we associate the vector

$$\bar{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_n, 0)$$

of ${V_p}^{n+1}$. Thus we have a subset $\bar{A}$ of ${V_p}^{n+1}$ containing s^k elements and also possessing the property that $d(\bar{\alpha},\bar{\alpha}') \geqq 2t + 1$; $\bar{\alpha},\bar{\alpha}' \in \bar{A}$, $\bar{\alpha} \neq \bar{\alpha}'$. Hence, by Theorem 1, we can obtain a reliable system design of order t and redundancy $n + 1 - k = r + 1$. It follows that

$$n_t(r + 1) \geqq k + r + 1 = n + 1 = n_t(r) + 1.$$

Theorem 3: If for a reliable system design of order t and redundancy r for k Boolean p-functions we have

$$n_t(r - 1) - (r - 1) < k \leqq n_t(r) - r, \tag{9}$$

then the design is minimally redundant.

Proof: If possible, suppose the system is not minimally redundant. Then there exists a reliable system design of order t and redundancy

$r - c$ for k Boolean p-functions where c is some positive integer. Then we have $n_t(r - c) \geqq k + (r - c)$. By Lemma 1,

$$\begin{aligned} n_t(r - 1) &\geqq n_t(r - c) + (c - 1) \\ &\geqq k + (r - 1). \end{aligned} \tag{10}$$

The inequality (10) contradicts the inequality (9); hence the theorem is established.

IV. LINEAR SYSTEM DESIGNS

In this section we shall consider a particular subclass of system designs called the *linear system designs*. To define the linear system designs, we have to use the theory of finite fields. Let K be the finite field of characteristic 2 containing $s = 2^p$ elements and x denote a primitive element of K. Any binary p-tuple $(a_0, a_1, \cdots, a_{p-1})$ will be made to correspond to the element $a_0 + a_1x + \cdots + a_{p-1}x^{p+1}$ of K and vice versa. An element $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_n)$ of V_p^n now will be considered as an n-vector with elements in K. The weight $\omega(\alpha)$ of α is equal to the number of nonnull elements among $\alpha_1, \alpha_2, \cdots, \alpha_n$. The sum of two vectors $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_n)$ and $\alpha' = (\alpha_1', \alpha_2', \cdots, \alpha_n')$ is defined to be

$$\alpha + \alpha' = (\alpha_1 + \alpha_1', \alpha_2 + \alpha_2', \cdots, \alpha_n + \alpha_n').$$

Obviously V_p^n is a vector space over K. Consider a system design for k Boolean p-functions. Suppose the encoder function is

$$\varphi = (\varphi_1, \varphi_2, \cdots, \varphi_n).$$

For every value X' of the input variable X,

$$\varphi(X') = [\varphi_1(X'), \varphi_2(X'), \cdots, \varphi_n(X')]$$

is a vector belonging to V_p^n. Let

$$A = \{\varphi(X') \mid X' \in B_m\}. \tag{11}$$

Definition 3: A system design for k Boolean p-functions is said to be a *linear system design* if the subset A of V_p^n defined by (11) is a vector space over K.

Lemma 2: A necessary and sufficient condition that a reliable linear system design for k Boolean p-functions is of order t is that the weight of any nonnull vector of the set A defined in (11) is not less than $(2t + 1)$.

Proof: Because of Theorem 1, it would be sufficient to show that

$$d(\alpha,\alpha') \geqq 2t + 1; \qquad \alpha,\alpha' \in A, \quad \alpha \neq \alpha'. \tag{12}$$

By definition $d(\alpha,\alpha') = \omega(\alpha + \alpha')$. Since A is a vector space, $\alpha + \alpha'$ is an element of A and also $\alpha + \alpha'$ is a nonnull element of A. Hence it follows that (12) will hold if and only if $\omega(\alpha) \geqq 2t + 1$ for every nonnull element α of A.

Definition 4: A matrix M with elements in K will be said to have the (P_t)-property if no t rows of the matrix are linearly dependent.

Theorem 4: A necessary and sufficient condition for the existence of a reliable linear system design of order t and redundancy r for k Boolean p-functions is that there exists a matrix M with $n = (k + r)$ rows and r columns with elements in K which possesses (P_{2t})-property.

Proof: Sufficiency. Suppose the matrix M is given by

$$M = \begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1r} \\ m_{21} & m_{22} & \cdots & m_{2r} \\ \vdots & \vdots & & \vdots \\ m_{n1} & m_{n2} & & m_{nr} \end{bmatrix}. \tag{13}$$

Let A denote the vector space orthogonal to the vector space generated by the r rolumn vectors of M. A contains at least s^k elements. It would be sufficient to show that the weight of any nonnull vector α of A is at least $(2t + 1)$. If possible, suppose A contains a nonnull vector with weight less than $(2t + 1)$. For simplicity of writing assume that the vector $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_{2t}, 0, \cdots, 0)$ belongs to A where $\alpha_1, \alpha_2, \cdots, \alpha_{2t}$ are nonzero elements of K. Then we have

$$\alpha_1 m_{1i} + \alpha_2 m_{2i} + \cdots + \alpha_{2t} m_{2ti} = 0; \qquad i = 1,2,\cdots,r. \tag{14}$$

Equation (14) implies that the first $2t$ vectors of the matrix M are linearly dependent which is a contradiction. This completes the proof of sufficiency. Necessity can be proved by exactly similar arguments.

The reader acquainted with the literature on error-correcting linear codes would recognize from Theorem 4 that a reliable linear system design of order t for k Boolean p-functions exists if and only if a t-error correcting linear code in $s(= 2^p)$ symbols with n places and k information places exists. Lemma 1 and Theorem 4 given above are not new results; they were proved by Bose[7] and Zierler[8] in a different form. We have included short proofs for these results for the sake of completeness.

V. MINIMALLY REDUNDANT LINEAR SYSTEM DESIGNS OF ORDER 1

In this section we shall give methods for constructing minimally redundant linear system designs of order 1 for k Boolean p-functions for any arbitrary value of k and p.

Theorem 5:

$$n_1(r) = \frac{s^r - 1}{s - 1}.$$

Proof: In Section III we proved that

$$n_1(r) \leqq \frac{s^r - 1}{s - 1}.$$

Hence it would be sufficient to show that

$$n_1(r) \geqq \frac{s^r - 1}{s - 1}. \tag{15}$$

To prove (15) we shall construct a matrix M with r columns and $n = (s^r - 1)/(s - 1)$ rows which has (P_2)-property. We shall denote the elements of K by $0,1,\alpha_2, \cdots, \alpha_{s-1}$, where 0 is the null element and 1 is the multiplicative identity. Consider the matrix M given by

$$M = \begin{bmatrix} M_1 \\ I_r \end{bmatrix}, \tag{16}$$

where I_r is the identity matrix with r rows and r columns and M_1 is a matrix with r columns and k $(= n - r)$ rows given below:

$$M_1 = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 & 1 \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 & \alpha_{s-1} \\ 0 & 0 & \cdots & 1 & 1 & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & \alpha_{s-1} & \alpha_{s-1} \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ 1 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ 1 & \alpha_{s-1} & & \alpha_{s-1} & \alpha_{s-1} & \alpha_{s-1} \end{bmatrix}. \tag{17}$$

It can be easily checked that the matrix M has (P_2)-property; i.e., no two rows of M are linearly dependent. This completes the proof of Theorem 5.

It should be observed that Theorem 5 enables us to construct minimally redundant system designs of order 1 for any arbitrary values of k and p. For given k, we find out the integer r for which

$$n_1(r - 1) - (r - 1) < k \leqq n_1(r) - r.$$

If $n_1(r) - r = k$, we construct the matrix M with r columns and $n_1(r)$ rows as defined in (16) and then obtain the system design as illustrated

in the proof of Theorem 4. If $n_1(r) - r > k$, then we delete $n_1(r) - (k + r)$ rows from M_1 and thus obtain a matrix M with (P_2)-property which has r columns and $(k + r)$ rows. From Theorem 3, is follows that the resulting system design will be minimally redundant. Now we shall give explicitly the encoder function of the minimally redundant system designs of order 1. Let

$$\underset{(n\times k)}{B} = \begin{bmatrix} I_k \\ M_1' \end{bmatrix},$$

where I_k is the identity matrix with k rows and k columns and M_1' is the transpose of the matrix M_1. It can be verified that the k column vectors of B are orthogonal to each of the r column vectors of M. The k column vectors of B generate the vector space A and every nonnull vector of A has weight greater than $2t$. For the sake of convenience of description, we write

$$M_1' = \begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1k} \\ \vdots & \vdots & & \vdots \\ m_{r1} & m_{r2} & \cdots & m_{rk} \end{bmatrix}.$$

To define the encoder function φ, we must set up a one-to-one correspondence between the s^k vectors of $V_p^{\,k}$ and the s^k vectors of A. We make the vector $(\alpha_1, \alpha_2, \cdots, \alpha_k)$ of $V_p^{\,k}$ correspond to the vector $(\alpha_1, \alpha_2, \cdots, \alpha_k, \alpha_{k+1}, \cdots, \alpha_n)$ of A, where

$$\begin{aligned} \alpha_{k+1} &= \alpha_1 m_{11} + \alpha_2 m_{12} + \cdots + \alpha_k m_{1k} \\ &\vdots \qquad\quad \vdots \qquad\quad \vdots \qquad\qquad \vdots \\ \alpha_n &= \alpha_1 m_{r1} + \alpha_2 m_{r2} + \cdots + \alpha_k m_{rk}. \end{aligned}$$

Hence, if

$$\begin{aligned} f(X') &= [f_1(X'), f_2(X'), \cdots, f_k(X')] \\ &= (\alpha_1, \alpha_2, \cdots, \alpha_k), \\ \varphi(X') &= [\varphi_1(X'), \varphi_2(X'), \cdots, \varphi_k(X'), \varphi_{k+1}(X'), \cdots, \varphi_n(X')] \\ &= (\alpha_1, \alpha_2, \cdots, \alpha_k, \alpha_{k+1}, \cdots, \alpha_n). \end{aligned}$$

Therefore it follows that we have

$$\begin{aligned} \varphi_1(X) &= f_1(X) \\ &\vdots \qquad \vdots \\ \varphi_k(X) &= f_k(X), \\ \varphi_{k+1}(X) &= m_{11} f_1(X) + m_{12} f_2(X) + \cdots + m_{1k} f_k(X) \\ &\vdots \qquad\qquad \vdots \qquad\qquad\quad \vdots \qquad\qquad\quad \vdots \\ \varphi_n(X) &= m_{r1} f_1(X) + m_{r2} f_2(X) + \cdots + m_{rk} f_k(X). \end{aligned} \tag{18}$$

It should be pointed out that in (18), for any particular value X' of X, $f_k(X')$ and m_{ji} $(j = 1,2,\cdots,r;\ i = 1,2,\cdots,k)$ are elements of K and the operations of addition and multiplication are as in K. The corrector function can be built up, from the vector space A by the method already described in connection with the proof of the sufficiency part of Theorem 1. In the following, we shall give an alternative simpler method of building up a corrector subsystem which will correct one error in the encoder subsystem.

Suppose the output of the encoder subsystem is the vector $\gamma = (\gamma_1, \gamma_2, \cdots, \gamma_n)$ of V_p^n. If X' is the value of the input variable, then $\gamma = \varphi(X') + \epsilon$ where $\varphi(X') = (\varphi_1(X'), \varphi_2(X'), \cdots, \varphi_n(X'))$ and $\epsilon = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)$ is the error vector. From our method of constructing the encoder function, we have

$$\underset{(1\times n)}{\varphi(X')}\ \underset{(n\times r)}{M} = \underset{(1\times r)}{\mathbf{0}}.$$

Hence

$$\gamma M = \epsilon M. \tag{19}$$

Let

$$\epsilon M = \delta = (\delta_1, \delta_2, \cdots, \delta_n).$$

Then

$$\begin{aligned} \delta_1 &= \gamma_{k+1} + m_{11}\gamma_1 + \cdots + m_{1k}\gamma_k \\ &\vdots \quad\quad \vdots \quad\quad\quad \vdots \\ \delta_r &= \gamma_n + m_{r1}\gamma_1 + \cdots + m_{rk}\gamma_k. \end{aligned} \tag{20}$$

If there is error in one block of the encoder subsystem, ϵ will have one nonzero element among its coordinates. Suppose the lth coordinate of ϵ is a nonzero element λ of K. Then $\epsilon = (0,0,\cdots,\lambda,0,\cdots,0)$, $1 \leqq l \leqq n$. Let us denote the lth row vector of the matrix M by

$$R_l = (c_{l1}, c_{l2}, \cdots, c_{lr}).$$

Then the following equations hold:

$$\begin{aligned} \delta &= \epsilon M = \lambda R_l \\ \delta_j &= \lambda c_{lj}, \qquad j = 1,2,\cdots,r. \end{aligned} \tag{21}$$

Conversely, if for a given output vector γ the vector δ computed by (20) satisfies (21) and there is one block of the encoder subsystem in error, then the error vector ϵ will have λ as its lth coordinate and zero as the other coordinates.

From the above discussion, it is clear that a corrector subsystem which

observes the rules given below will accomplish the job of correcting errors in one block of the encoder subsystem and produce the k Boolean p-function $f_1(X), f_2(X), \cdots, f_k(X)$ as its output. The rules are

i. Compute δ as defined in (20).

ii. If δ is the null vector, the k outputs would be given by $\beta_i = \gamma_i$, $i = 1,2,\cdots,k$.

iii. If δ is not the null vector, find out the integer l for which the vector λR_l for some $\lambda \in K$ has maximum number of common coordinates with δ. If $l > k$, the k outputs are $\beta_i = \gamma_i$, $i = 1,2,\cdots,k$. If $l < k$, the k outputs are $\beta_1 = \gamma_i$, $\beta_2 = \gamma_2$, $\cdots$, $\beta_l = \gamma_l + \lambda, \cdots, \beta_k = \gamma_k$.

VI. AN EXAMPLE

In this section we shall give an example to show how the theory developed in this paper can be applied.

Suppose $m = 3$, $p = 2$, $k = 3$ and $t = 1$. From Section V, we can see that for the minimally redundant system $r = 2$. Suppose the three Boolean two-functions to be synthesized are

$$\begin{aligned} f_1(X) &= [f_{11}(X), f_{12}(X)] \\ &= (X_1 \cdot X_2, X_1 \oplus X_2), \\ f_2(X) &= [f_{21}(X), f_{22}(X)] \\ &= (X_2 \cdot X_3, X_2 \oplus X_3), \\ f_3(X) &= [f_{31}(X), f_{32}(X)] \\ &= (X_1 \cdot X_3, X_1 \oplus X_3) \end{aligned}$$

where the symbols $\oplus$ and $\cdot$ are respectively used to denote the Boolean operations of additions (OR) and multiplication (AND) between two binary variables. Let k denote the field containing four elements. Let t be a primitive element of the field. The polynomial $t^2 + t + 1$ is a minimum function and every element of the field satisfies the equation $x^3 = 1$. The four elements are shown below in terms of the primitive element t, and their correspondence with binary 2-vectors is also pointed out:

$$\begin{aligned} \alpha_0 &= 0 = 0 + 0t \leftrightarrow (0,0), \\ \alpha_1 &= 1 = 1 + 0t \leftrightarrow (1,0), \\ \alpha_2 &= t = 0 + 1t \leftrightarrow (0,1), \\ \alpha_3 &= t^2 = 1 + 1t \leftrightarrow (1,1). \end{aligned}$$

In view of the correspondence between the binary 2-vectors and the elements of K, any particular value of a Boolean 2-function will be considered as an element of K. For example, if

$$f_1(X') = (1,1),$$

then

$$f_1(X') = \alpha_3 .$$

Addition and multiplication between the elements of K are shown in the tables given below:

Addition Table

	α_0	α_1	α_2	α_3
α_0	α_0	α_1	α_2	α_3
α_1	α_1	α_0	α_3	α_2
α_2	α_2	α_3	α_0	α_1
α_3	α_3	α_2	α_1	α_0

Multiplication Table

	α_0	α_1	α_2	α_3
α_0	α_0	α_0	α_0	α_0
α_1	α_0	α_1	α_2	α_3
α_2	α_0	α_2	α_3	α_1
α_3	α_0	α_3	α_1	α_2

The sum of two elements α_i and α_j is obtained by adding the corresponding polynomials in t modulo 2 ($i, j = 0, 1, 2, 3$). The product of two elements is obtained by multiplying the corresponding polynomials modulo 2 and modulo $(t^2 + t + 1)$. From Section V we have $m_{11} = m_{12} = m_{13} = \alpha_1$ and $m_{21} = \alpha_1$, $m_{22} = \alpha_2$ and $m_{23} = \alpha_3$. Therefore the five encoder Boolean 2-functions are given by

$$\varphi_i(X') = f_i(X'), \qquad i = 1, 2, 3,$$

$$\varphi_4(X') = f_1(X') + f_2(X') + f_3(X')$$

and

$$\varphi_5(X') = f_1(X') + \alpha_2 f_2(X') + \alpha_3 f_3(X').$$

Hence, $\varphi_{41}(X) = (X_1 \cdot X_2) + (X_2 \cdot X_3) + (X_1 \cdot X_3)$ and $\varphi_{42}(X) = (X_1 \oplus X_2) + (X_2 \oplus X_3) + (X_3 \oplus X_1)$, where $+$, $\oplus$ and $\cdot$ respectively denote mod 2 addition, Boolean addition (OR) and Boolean multiplication (AND) between two binary variables.

The truth table for the two Boolean functions φ_{51} and φ_{52} is given below:

X_1	X_2	X_3	φ_{51}	φ_{52}
0	0	0	0	0
0	0	1	0	1
0	1	0	1	0
0	1	1	0	1
1	0	0	1	0
1	0	1	1	1
1	1	0	1	0
1	1	1	0	0

The computation of this table will be illustrated by one example. Suppose $X_1 = 0$, $X_2 = 1$, $X_3 = 1$. Then $f_1(X) = (0,1) = \alpha_2$, $f_2(X) = (1,1) = \alpha_3$ and $f_3(X) = (0,1) = \alpha_2$. So $\varphi_5(X) = \alpha_2 + \alpha_2\alpha_3 + \alpha_3\alpha_2 = \alpha_2 = (0,1)$. And so $\varphi_{51}(X) = 0$ and $\varphi_{52}(X) = 1$. The corrector subsystem uses the outputs $\gamma_i = (\gamma_{i1},\gamma_{i2})$, $i = 1, 2, 3, 4, 5$, of the encoder subsystem as inputs. The final outputs $\beta_i = (\beta_{i1}, \beta_{i2})$, $i = 1, 2, 3$ of the corrector subsystem will be built up in several stages. From Section V, $\delta_1 = \gamma_4 + \gamma_1 + \gamma_2 + \gamma_3$ and $\delta_2 = \gamma_5 + \gamma_1 + \alpha_2\gamma_2 + \alpha_3\gamma_3$. At the first stage the corrector subsystem synthesizes $\eta_i = \alpha_i\gamma_1$, $i = 2, 3$. The truth tables for (η_{i1}, η_{i2}), $i = 2, 3$ are given below:

γ_{21}	γ_{22}	η_{21}	η_{22}	γ_{31}	γ_{32}	η_{31}	η_{32}
0	0	0	0	0	0	0	0
0	1	1	1	0	1	1	0
1	0	0	1	1	0	1	1
1	1	1	0	1	1	0	1

At the second stage, the binary 2-tuples δ_1, and δ_2 are synthesized. We have

$$\begin{aligned}
\delta_{11} &= \gamma_{41} + \gamma_{11} + \gamma_{21} + \gamma_{31}, \\
\delta_{21} &= \gamma_{42} + \gamma_{21} + \gamma_{22} + \gamma_{32}, \\
\delta_{21} &= \gamma_{51} + \gamma_{11} + \eta_{21} + \eta_{31}, \\
\delta_{22} &= \gamma_{52} + \gamma_{12} + \eta_{22} + \eta_{32}.
\end{aligned} \tag{22}$$

The addition between the binary variables in (22) is modulo 2 addition. At the third stage, the three binary 2-tuples, ϵ_1, ϵ_2 and ϵ_3, which are the first three coordinates of the error vector ϵ are synthesized as Boolean functions of the δ's. The part of the truth table in which at least one of the ϵ's takes the value 1 is given below:

δ_{11}	δ_{12}	δ_{21}	δ_{22}	ϵ_{11}	ϵ_{12}	ϵ_{21}	ϵ_{22}	ϵ_{31}	ϵ_{32}
0	1	0	1	0	1	0	0	0	0
0	1	1	0	0	0	0	0	0	1
0	1	1	1	0	0	0	1	0	0
1	0	0	1	0	0	1	0	0	0
1	0	1	0	1	0	0	0	0	0
1	0	1	1	0	0	0	0	1	0
1	1	0	1	0	0	0	0	1	1
1	1	1	0	0	0	1	1	0	0
1	1	1	1	1	1	0	0	0	0

The truth table given below is computed from the rules given in Section V. In the case of our example,

$$M = \begin{bmatrix} \alpha_1 & \alpha_1 \\ \alpha_1 & \alpha_2 \\ \alpha_1 & \alpha_3 \\ \alpha_1 & \alpha_0 \\ \alpha_0 & \alpha_1 \end{bmatrix}.$$

Suppose $\delta_{11} = 0$, $\delta_{12} = 1$, $\delta_{21} = 1$ and $\delta_{22} = 1$. Then $\delta_1 = \alpha_2$ and $\delta_2 = \alpha_3$. Since $\delta_2 = \alpha_2\alpha_1$ and $\delta_2 = \alpha_2^2$, the vector δ is a scalar multiple of the second row vector of M. Therefore it follows that $\epsilon_1 = \alpha_0$, $\epsilon_2 = \alpha_2$ and $\epsilon_3 = \alpha_0$. Hence, $\epsilon_{11} = 0$, $\epsilon_{12} = 0$, $\epsilon_{21} = 0$, $\epsilon_{22} = 1$, $\epsilon_{31} = 0$ and $\epsilon_{32} = 1$.

In the example considered above the number of input variables was small and the Boolean functions required to be synthesized were chosen to be very simple. Therefore the corrector subsystem would probably require more equipment than the encoder subsystem. However, it should be noted that the design of the corrector subsystem is independent of the number of binary input variables and the nature of the original Boolean functions. This design depends only on p and k. Therefore when the number of input variables is large and the Boolean functions required to be synthesized are complicated, the amount of equipment required for the corrector subsystem may be small in comparison to that required for the whole system. This is very desirable, since we assume that the corrector subsystem is highly reliable. The example shows how we can build up the logical design of the corrector subsystem in any general case. However, the author believes that it is possible to build up much more economical corrector subsystems using sequential circuits. Of course, one then pays the penalty of taking a

longer time to correct the errors. Such economical corrector subsystems are discussed in the companion paper,[1] in which minimally redundant reliable systems which correct faults of more than one block are also given.

VII. ACKNOWLEDGMENTS

The author wishes to thank T. H. Crowley, D. B. Armstrong, J. P. Runyon and B. A. Tague for many stimulating and useful discussions.

REFERENCES

1. Armstrong, D. B., this issue, p. 577.
2. Moore, E. F. and Shannon, C. E., Reliable Circuits Using Less Reliable Relays, J. Frank. Inst., **262,** 1956, pp. 191; 281.
3. Tryon, J. G., Redundant Logic Circuitry, U. S. Patent No. 2,942,193.
4. Von Neumann, J., Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components, *Automata Studies*, Annals of Math. Studies, No. 34, Princeton Univ. Press, Princeton, N. J., 1956, pp. 44–98.
5. Lofgren, L., Automata of High Complexity and Methods of Increasing Their Reliability by Redundancy, Inf. & Cont., **1,** 1958, p. 127.
6. Hamming, R. W., Error Detecting and Error Correcting Codes, B.S.T.J., **29,** 1950, p. 147.
7. Bose, R. C., Mathematical Theory of the Symmetrical Factorial Design, Sankhya, **8,** 1947, p. 155.
8. Zierler, N., A Class of Cyclic, Linear, Error-Correcting Codes in p^m Symbols, Group Report 55-19, Lincoln Laboratory.

Mode Conversion in Metallic and Helix Waveguide*

By H. G. UNGER

(Manuscript received September 29, 1960)

Helix waveguide, composed of closely wound insulated copper wire covered with an absorptive or reactive jacket, transmits circular electric waves with low loss. Mechanical imperfections, such as curvature and deformation, cause mode conversion and degrade the transmission. Generalized telegraphist's equations describe propagation in an imperfect helix waveguide with coupling between the many propagating waves. The coefficients of coupling depend strongly on the outside jacket. However, the sum of the squares of the coupling coefficients is independent of the jacket for circular electric waves. As a consequence, the average circular electric wave loss in a helix waveguide with random imperfections is also nearly independent of the jacket and the same as in a metallic pipe.

I. INTRODUCTION

Helix waveguide consisting of closely wound insulated copper wire covered with an electrically absorptive or reactive jacket (Fig. 1) is a good transmission medium for circular electric waves.[1] In long distance communication with waveguides it is useful as a mode filter, for negotiating bends or particularly as the transmission line proper instead of a metallic waveguide.[2]

The loss of circular electric waves decreases steadily with frequency only in a perfect metallic waveguide.[3,4] A similar situation prevails for helix waveguide. Any deviations from a round and straight guide will add to the transmission loss. At such imperfections power is converted from the circular electric wave into other propagating modes and reconverted. The mode conversion-reconversion effects increase the loss and degrade the transmission characteristics.

* Parts of this paper were presented at the I.R.E. Professional Group on Microwave Theory and Techniques Symposium, San Diego, California, 1960.

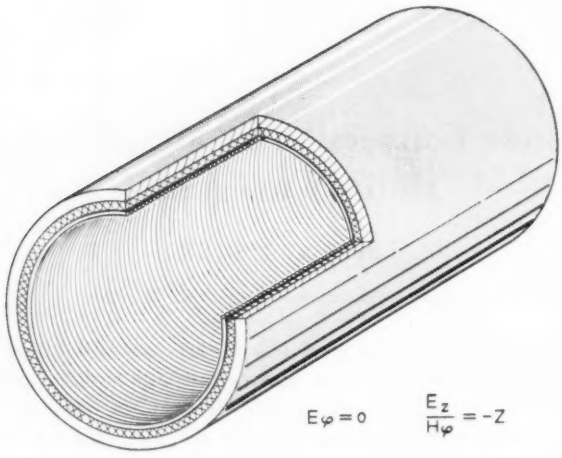


Fig. 1 — Helix waveguide and boundary conditions.

II. GENERALIZED TELEGRAPHIST'S EQUATIONS AND MATRIX REPRESENTATION

Wave propagation in imperfect helix waveguide is most conveniently described with generalized telegraphist's equations.[5] This is a representation in terms of the normal modes of the perfect waveguide. The solution of the perfect waveguide is perturbed in the imperfect waveguide. Physically the perturbation appears as coupling between the normal modes of the perfect waveguide.

In the perfect metallic waveguide, wave propagation is completely described by a set of independent first-order differential equations

$$\frac{dA_n}{dz} = -\kappa_{nn} A_n , \tag{1}$$

where A_n are the amplitudes of the TE_{pn} or TM_{pn} modes normalized with respect to power and κ_{nn} are their propagation constants. z is the axial coordinate. A square matrix with only the diagonal terms $\kappa_{nn} = jh_n$,

$$K = \begin{bmatrix} jh_0 & 0 & 0 & \cdots \\ 0 & jh_1 & 0 & \cdots \\ 0 & 0 & jh_2 & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix}, \tag{2}$$

describes the perfect metallic waveguide completely.

The perfect helix waveguide may be considered a perturbed metallic wavèguide. The boundary conditions for the tangential electric field

$$E_\varphi = 0 \qquad \text{and} \qquad E_z = 0 \tag{3}$$

are replaced by

$$E_\varphi = 0 \qquad \text{and} \qquad \frac{E_z}{H_\varphi} = -Z, \tag{4}$$

where Z is the wall impedance that the jacket of the helix waveguide presents at the helix interface.

To describe wave propagation in the helix waveguide with metallic waveguide modes, off-diagonal terms have to be added in the diagonal matrix of the metallic waveguide and the diagonal terms are perturbed:

$$K = \begin{bmatrix} \kappa_{00} & 0 & 0 & 0 \\ 0 & \kappa_{11} & \kappa_{12} & \cdots \\ 0 & \kappa_{12} & \kappa_{22} & \cdots \\ 0 & \vdots & \vdots & \end{bmatrix}. \tag{5}$$

Only the propagation constant $\kappa_{00} = jh_0$ of the circular electric wave stays unperturbed, and its off-diagonal terms remain zero. In other words, to describe wave propagation in helix waveguides with normal modes of metallic waveguide, coupling has to be introduced between these modes, and their propagation constants are modified. Only circular electric waves stay unchanged[1].

It is now expedient to transform from the normal modes of metallic waveguide A_n to the normal modes of helix waveguide E_n :

$$A = LE. \tag{6}$$

The so-called modal matrix L of K transforms K to its diagonal form Γ:

$$L^{-1}KL = \Gamma. \tag{7}$$

Since K is symmetrical the modal matrix L is orthonormal. It obeys:

$$L_t = L^{-1}. \tag{8}$$

Its transpose is equal to its inverse. The diagonal terms γ_{nn} of Γ are the propagation constants of normal modes in helix waveguide. The circular electric wave remains unaffected by this transformation:

$$\gamma_{00} = \kappa_{00} = jh_0 .$$

It is now possible to consider an imperfect helix waveguide, perturbed by curvature or cross-sectional deformations in the same terms. The matrix K' then has off-diagonal elements also in the row and column associated with the circular electric wave:

$$K' = \begin{bmatrix} \kappa_{00} & \kappa_{01} & \kappa_{02} & \cdots \\ \kappa_{01} & \kappa_{11} & \kappa_{12} & \cdots \\ \kappa_{02} & \kappa_{12} & \kappa_{22} & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix}. \tag{9}$$

These new off-diagonal elements represent coupling between the circular electric wave and other modes in an imperfect helix waveguide. They are the coupling coefficients of generalized telegraphist's equations. When Maxwell's equations for the imperfect helix waveguide are converted into generalized telegraphist's equations, it is found that the coupling coefficients κ_{0n} in K' for curvature and cross-sectional deformation are independent of the wall impedance and the same as in metallic waveguide. Generalized telegraphist's equations for deformed helix waveguide are found in the Appendix.

The independence of the κ_{0n}'s from the wall impedance has an important bearing on the coupling coefficients between the normal modes in helix waveguide. If the new K' for the imperfect helix waveguide is transformed by the previous modal matrix L, there results

$$L_t K' L = \Gamma'. \tag{10}$$

The new matrix Γ' has off-diagonal elements c_{0n} which are caused by the imperfection. They are the coefficients of coupling between the normal modes of the perfect helix waveguide. In terms of the elements of K' and L they are:

$$c_{0n} = \sum_{m=1}^{\infty} \kappa_{0m} l_{mn} . \tag{11}$$

If this expression is squared and summed over all n, then, since the l's are elements of an orthonormal matrix, the resulting expression does not contain any l's:

$$\sum_n c_{0n}^2 = \sum_m \kappa_{0m}^2 . \tag{12}$$

It is recalled that the κ_{0m} are coupling coefficients between normal modes in an imperfect metallic waveguide. Consequently, the sum of the squares of the coupling coefficients c_{0n} in an imperfect helix waveguide is independent of the wall impedance and the same as in metallic waveguide.

III. AVERAGE MODE CONVERSION LOSS FROM RANDOM IMPERFECTIONS

The following statement will be proved: The average mode conversion loss of certain kinds of random imperfections in helix waveguide depends only on the sum of the square of all coupling coefficients.

To determine mode conversion, first of all, generalized telegraphist's equations have to be solved. With the elements of the perturbed Γ matrix the coupled line equations are

$$\frac{dE_n}{dz} = -\gamma_n E_n - \sum_m c_{nm} E_m . \tag{13}$$

When only a circular electric wave of unit amplitude, $E_0(0) = 1$, is incident and the imperfections are small, a first-order solution at $z = L$ is:

$$| E_0(L) | = \left| 1 - \sum_n \int_0^L e^{(\gamma_0 - \gamma_n)u}\, du \int_0^{L-u} c_{0n}(s) c_{0n}(s + u)\, ds \right| . \tag{14}$$

The coupling coefficients are proportional to the geometric imperfection δ:

$$c_{0n}(z) = C_{0n}\delta(z). \tag{15}$$

Let the imperfections be a stationary random process with covariance

$$\rho(u) = \langle \delta(z)\delta(z + u) \rangle \tag{16}$$

and spectral distribution

$$S(\xi) = \int_{-\infty}^{+\infty} \rho(u)\, e^{-j2\pi\xi u}\, du, \tag{17}$$

where ξ is the spatial frequency of the geometric imperfection. Then, as Rowe first pointed out,[6] the average output amplitude $\langle | E_0(L) | \rangle$ can be expressed in terms of the covariance $\rho(u)$:

$$\langle | E_0(L) | \rangle = \left| 1 - \sum_n {C_{0n}}^2 \int_0^L (L - u)\, e^{(\gamma_0 - \gamma_n)u} \rho(u)\, du \right| .$$

If, furthermore, the correlation between imperfections any appreciable distance apart is small the covariance drops off very rapidly with increasing argument. Then in the expression for $\langle | E_0(L) | \rangle$ the exponential and the factor $(L - u)$ are constant for any weight of $\rho(u)$, and one obtains:

$$\langle | E_0(L) | \rangle = \left| 1 - L \int_0^\infty \rho(u)\, du \sum_n {C_{0n}}^2 \right| . \tag{18}$$

The average mode conversion loss $\langle \alpha \rangle$ can now be written in terms of

the spectral density S of the random imperfections δ and the coupling factors C_{0n} :

$$\langle \alpha \rangle = \tfrac{1}{2} S(0) \text{ Re } [\sum_n C_{0n}{}^2]. \tag{19}$$

With (12) and (15), it may be concluded from (19) that the average loss for circular electric waves in an imperfect helix waveguide is independent of the wall impedance and the same as in metallic waveguide that has the same geometrical imperfections.

The above derivation has assumed the covariance to drop off fast or the correlation distance to be small. This is the case for any imperfection created in the manufacturing process of the waveguide. Any manufacturing imperfections some reasonable distance apart are hardly correlated to each other. The effects of manufacturing imperfections for helix waveguide are therefore the same as in metallic waveguide. It is relatively easy to determine tolerances for metallic waveguide.[4] The above rule lets these tolerances be valid for helix waveguide and the very involved calculations for helix waveguide are not necessary.

Before accepting this rule the range of correlation distance for which it is valid must be examined. As a typical example, the covariance has been assumed exponential:

$$\rho(u) = \langle \delta^2 \rangle e^{-2\pi(|u|/L_0)}. \tag{20}$$

The average TE_{01} loss at 55 kmc has then been calculated for various helix waveguides of 2-inch inside diameter as a function of the correlation distance L_0.[7] Fig. 2 shows for deformed helix waveguide the rms of elliptical diameter differences which increase the TE_{01} loss by 10 per cent of the loss in a perfect copper pipe. Up to a correlation distance of one foot the curves almost coincide and indicate independence of the wall impedance.

For a curved helix waveguide the range is even larger. Fig. 3 shows for a curved helix waveguide the rms curvature under the same conditions. Random curvature of up to a 10-foot correlation distance causes nearly the same average TE_{01} loss in helix waveguide and in metallic waveguide.

For a correlation distance larger than 10 feet there is an ever growing dependence of curvature loss on wall impedance. But such curvature distributions do not occur in the manufacturing process. They are, however, representative of laying tolerances when the waveguide is installed with long bows to follow right of ways or the contour of the landscape. A properly designed helix waveguide can tolerate much more laying curvature than can metallic waveguide.

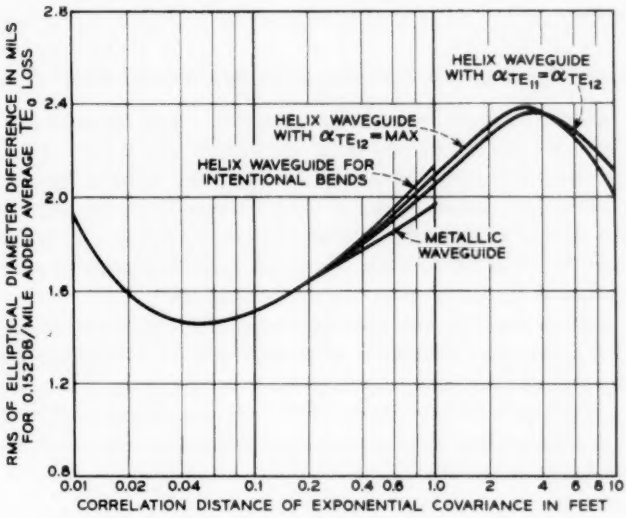


Fig. 2 — TE_{01} loss in round waveguide with random ellipticity, 2-inch inside diameter, at 55.5 kmc.

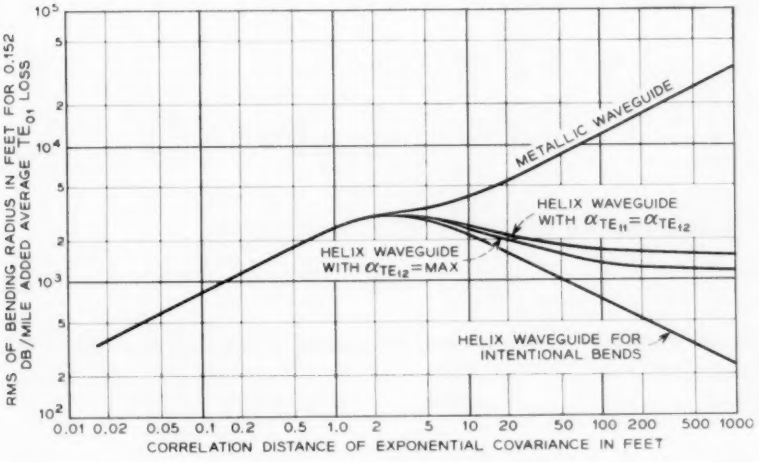


Fig. 3 — TE_{01} loss in round waveguide with random curvature, 2-inch inside diameter, at 55.5 kmc.

APPENDIX

Generalized Telegraphist's Equation for Noncylindrical Helix Waveguide

Maxwell's equations have been converted into generalized telegraphist's equations for curved helix waveguide elsewhere.[2] They have been represented in terms of normal modes of the metallic waveguide as well as in terms of normal modes of helix waveguide. In the former representation the coefficients of curvature coupling between circular electric waves and the modes of metallic waveguide were independent of the wall impedance and the same as in metallic waveguide.

The same is true for any cross-sectional deformation in helix waveguide. To prove this, Maxwell's equations will be converted into generalized telegraphist's equations in terms of modes of metallic waveguide for the deformed helix waveguide. This representation is different from another analysis, where the equations are written in terms of the normal modes of helix waveguide.[8]

If the radius of the deformed guide is

$$a_1 = a(1 + \delta),$$

where a is the nominal radius and the deformation δ is for the moment assumed to be only a function of φ of cylindrical coordinates $(r\varphi z)$, then the boundary conditions at $r = a$, are

$$E_\varphi + E_r \frac{d\delta}{d\varphi} = 0, \tag{21}$$

$$E_z = \frac{-Z}{\sqrt{1 + \left(\frac{d\delta}{d\varphi}\right)^2}} \left(H_\varphi + H_r \frac{d\delta}{d\varphi}\right). \tag{22}$$

The deformation is assumed to be small and smooth:

$$\delta \ll 1 \qquad \text{and} \qquad \frac{d\delta}{d\varphi} \ll 1. \tag{23}$$

The fields at $r = a_1$ can by series expansion be written in terms of the fields at $r = a$. The boundary conditions then take the approximate form:

$$E_\varphi = -\frac{\partial E_\varphi}{\partial r} a\delta - E_r \frac{d\delta}{d\varphi}, \tag{24}$$

$$E_z = -\frac{\partial E_z}{\partial r} a\delta - Z\left(H_\varphi + \frac{\partial H_\varphi}{\partial_r} a\delta + H_r \frac{d\delta}{d\varphi}\right). \tag{25}$$

Maxwell's equations in cylindrical coordinates for exponential time dependence $\epsilon^{j\omega t}$ are

$$\frac{1}{r}\frac{\partial E_z}{\partial \varphi} - \frac{\partial E_\varphi}{\partial z} = -j\omega\mu H_r, \tag{26}$$

$$\frac{\partial E_r}{\partial z} - \frac{\partial E_z}{\partial r} = -j\omega\mu H_\varphi, \tag{27}$$

$$\frac{1}{r}\frac{\partial (rE_\varphi)}{\partial r} - \frac{1}{r}\frac{\partial E_r}{\partial \varphi} = -j\omega\mu H_z, \tag{28}$$

$$\frac{1}{r}\frac{\partial H_z}{\partial \varphi} - \frac{\partial H_\varphi}{\partial z} = j\omega\epsilon E_r, \tag{29}$$

$$\frac{\partial H_r}{\partial z} - \frac{\partial H_z}{\partial r} = j\omega\epsilon E_\varphi, \tag{30}$$

$$\frac{1}{r}\frac{\partial (rH_\varphi)}{\partial r} - \frac{1}{r}\frac{\partial H_r}{\partial \varphi} = j\omega\epsilon E_z, \tag{31}$$

where μ and ϵ are permeability and permittivity of the waveguide interior.

The electromagnetic field is derived from two sets of wave functions, $T_{(n)}$ for TM waves and $T_{[n]}$ for TE waves of metallic waveguide:

$$\begin{aligned}
E_r &= \sum_n \left[V_{(n)} \frac{\partial T_{(n)}}{\partial r} + V_{[n]} \frac{\partial T_{[n]}}{r\partial \varphi} \right], \\
E_\varphi &= \sum_n \left[V_{(n)} \frac{\partial T_{(n)}}{r\partial \varphi} - V_{[n]} \frac{\partial T_{[n]}}{\partial r} \right], \\
H_r &= \sum_n \left[-I_{(n)} \frac{\partial T_{(n)}}{r\partial \varphi} + I_{[n]} \frac{\partial T_{[n]}}{\partial r} \right], \\
H_\varphi &= \sum_n \left[I_{(n)} \frac{\partial T_{(n)}}{\partial r} + I_{[n]} \frac{\partial T_{[n]}}{r\partial \varphi} \right].
\end{aligned} \tag{32}$$

The transverse field distribution is described by $T(r,\varphi)$ while the voltage and current coefficients $V(z)$ and $I(z)$ are functions of the coordinate z.

The T-functions satisfy the wave equation

$$\frac{1}{r}\left[\frac{\partial}{\partial r}\left(r\frac{\partial T}{\partial r} \right) + \frac{\partial}{\partial \varphi}\left(\frac{\partial T}{r\partial \varphi} \right) \right] = -\chi^2 T, \tag{33}$$

where χ is a separation constant, which takes on discrete values for the various normal modes. The T-functions are normalized so that

$$\int_S (\text{grad } T)\,(\text{grad } T)\,dS = \int_S (\text{flux } T)\,(\text{flux } T)\,dS = \chi^2 \int_S T^2\,dS = 1, \tag{34}$$

where S is the nominal cross section of the guide and the gradient and flux vectors of T are defined by:

$$\begin{aligned} \text{grad}_r\, T &= \frac{\partial T}{\partial r}, & \text{grad}_\varphi\, T &= \frac{\partial T}{r\partial \varphi}, \\ \text{flux}_r\, T &= \frac{\partial T}{r\partial \varphi}, & \text{flux}_\varphi\, T &= -\frac{\partial T}{\partial r}. \end{aligned} \tag{35}$$

Various orthogonality relations exist among the T functions:

$$\int_S T_{(n)} T_{(m)}\,dS = \int_S T_{[n]} T_{[m]}\,dS = 0,$$

$$\int_S (\text{grad } T_{(n)})\,(\text{grad } T_{(m)})\,dS = \int_S (\text{flux } T_{(n)})\,(\text{flux } T_{(m)})\,dS = 0, \tag{36}$$

$$\int_S (\text{grad } T_{[n]})\,(\text{grad } T_{[m]})\,dS = \int_S (\text{flux } T_{[n]})\,(\text{flux } T_{[m]})\,dS = 0$$

if $m \neq n$, and

$$\begin{aligned} \int_S (\text{grad } T_{(n)})\,(\text{flux } T_{[m]})\,dS &= \int_S (\text{grad } T_{[n]})\,(\text{flux } T_{(m)})\,dS \\ &= \int_S (\text{grad } T_{(n)})(\text{flux } T_{(m)})\,dS = 0 \end{aligned} \tag{37}$$

for all m and n.

To transform Maxwell's equations into generalized telegraphist's equations the series expansions (32) are substituted for the field components in (26) through (31). Certain combinations of the equations are integrated over the nominal cross section, and advantage is taken of the orthogonality relations (36) and (37).

For example, adding $-\partial T_{(m)}/r\partial\varphi$ times (26) and $\partial T_{(m)}/\partial r$ times (27) and integrating over the nominal cross section:

$$\frac{dV_{(m)}}{dz} + j\omega\mu\, I_{(m)} = a \int_0^{2\pi} E_z \frac{\partial T_{(m)}}{\partial r}\bigg|_a d\varphi + \chi_{(m)}{}^2 \int_S E_z\, T_{(m)}\,dS. \tag{38}$$

In the first term on the right-hand side of (38) the boundary condition

(25) is substituted for E_z. In the second term (31) is substituted for E_z:

$$\frac{dV_{(m)}}{dz} + j\frac{h_{(m)}^2}{\omega\epsilon} I_{(m)} = -Z \sum_n \left[I_{(n)} a \int_0^{2\pi} (1-\delta) \frac{\partial T_{(n)}}{\partial r} \frac{\partial T_{(m)}}{\partial r}\bigg|_a d\varphi + I_{[n]} \int_0^{2\pi} (1-\delta) \frac{\partial T_{[n]}}{\partial \varphi} \frac{\partial T_{(m)}}{\partial r}\bigg|_a d\varphi \right] - a^2 \int_0^{2\pi} \delta \frac{\partial E_z}{\partial r} \frac{\partial T_{(m)}}{\partial r}\bigg|_a d\varphi, \tag{39}$$

where $h^2 = \omega^2\mu\epsilon - \chi^2$.

Add $\partial T_{[m]}/\partial r$ times (26) and $\partial T_{[m]}/r\partial\varphi$ times (27) and integrate over the nominal cross section:

$$\frac{dV_{[m]}}{dz} + j\omega\mu I_{[m]} = \int_0^{2\pi} E_z \frac{\partial T_{[m]}}{\partial \varphi}\bigg|_a d\varphi. \tag{40}$$

The boundary condition (25) is substituted for E_z:

$$\frac{dV_{[m]}}{dz} + j\omega\mu I_{[m]} = -a \int_0^{2\pi} \delta \frac{\partial E_z}{\partial r} \frac{\partial T_{[m]}}{\partial \varphi}\bigg|_a d\varphi - Z \sum_n \left[I_{(n)} \int_0^{2\pi} (1-\delta) \frac{\partial T_{(n)}}{\partial r} \frac{\partial T_{[m]}}{\partial \varphi}\bigg|_a d\varphi + I_{[n]} \int_0^{2\pi} \frac{1-\delta}{a} \frac{\partial T_{[n]}}{\partial \varphi} \frac{\partial T_{[m]}}{\partial \varphi}\bigg|_a d\varphi \right]. \tag{41}$$

Add $-\partial T_{(m)}/\partial r$ times (29) and $-\partial T_{(m)}/r\partial\varphi$ times (30) and integrate over the nominal cross section:

$$\frac{dI_{(m)}}{dz} + j\omega\epsilon V_{(m)} = 0. \tag{42}$$

Add $-\partial T_{[m]}/r\partial\varphi$ times (29) and $\partial T_{[m]}/\partial r$ times (30) and integrate over the nominal cross section:

$$\frac{dI_{[m]}}{dz} + j\omega\epsilon V_{[m]} = \chi_{[m]}^2 \int_S H_z T_{[m]} \, dS. \tag{43}$$

For the right-hand side of (43) integrate $T_{[m]}$ times (28) over the nominal cross section:

$$-j\omega\mu \int_S H_z T_{[m]} \, dS = V_{[m]} + a \int_0^{2\pi} E_\varphi T_{[m]}\bigg|_a d\varphi. \tag{44}$$

Substitute the boundary condition (24) for E_φ and perform the partial integration:

$$\int_0^{2\pi} E_r \frac{d\delta}{d\varphi} T_{[m]}\bigg|_a d\varphi = -\int_0^{2\pi} \delta \left(\frac{\partial E_r}{\partial \varphi} T_{[m]} + E_r \frac{\partial T_{[m]}}{\partial \varphi} \right)\bigg|_a d\varphi. \tag{45}$$

With (44) and (45), (43) reads:

$$\frac{dI_{[m]}}{dz} + j\frac{h_{[m]}^2}{\omega\mu} V_{[m]} = ja\frac{\chi_{[m]}^2}{\omega\mu}\left(\int_0^{2\pi} \delta E_r \frac{\partial T_{[m]}}{\partial\varphi}\bigg|_a d\varphi - a\sum_n V_{[n]}\chi_{[n]}^2 \int_0^{2\pi} \delta T_{[m]} T_{[n]}\bigg|_a d\varphi\right). \tag{46}$$

Equations (39), (41), (42) and (46) describe coupling between all modes. For the present consideration, all terms must be retained which contain the wall impedance Z, because in practical helix waveguide the wall impedance may be quite large. The deformation, however, is small, and to analyze circular electric wave transmission only δ terms need be retained that describe direct coupling between circular electric and other waves. The above equations then reduce to:

$$\frac{dV_{(m)}}{dz} + j\frac{h_{(m)}^2}{\omega\epsilon} I_{(m)} = -Z\sum_n \left(I_{(n)} a\int_0^{2\pi} \frac{\partial T_{(n)}}{\partial r}\frac{\partial T_{(m)}}{\partial r} d\varphi + I_{[n]}\int_0^{2\pi} \frac{\partial T_{[n]}}{\partial\varphi}\frac{\partial T_{(m)}}{\partial r} d\varphi\right), \tag{47}$$

$$\frac{dV_{[m]}}{dz} + j\omega\mu I_{[m]} = -Z\sum_n \left(I_{(n)}\int_0^{2\pi} \frac{\partial T_{(n)}}{\partial r}\frac{\partial T_{[m]}}{\partial\varphi} d\varphi + I_{[n]}\frac{1}{a}\int_0^{2\pi} \frac{\partial T_{[n]}}{\partial\varphi}\frac{\partial T_{[m]}}{\partial\varphi} d\varphi\right) \tag{48}$$

$$\frac{dI_{(m)}}{dz} + j\omega\epsilon V_{(m)} = 0 \tag{49}$$

$$\frac{dI_{[m]}}{dz} + j\frac{h_{[m]}^2}{\omega\mu} V_{[m]} = -j\frac{a^2}{\omega\mu}\sum_n V_{[n]}\chi_{[m]}^2\chi_{[n]}^2 \int_0^{2\pi} \delta\, T_{[m]}\, T_{[n]}\, d\varphi. \tag{50}$$

All the integrals are taken along the nominal circumference.

Alternatively, in terms of voltages and currents the equations are more conveniently written in terms of amplitudes A of forward- and B of backward-traveling waves. The mode current and voltage are related to the mode amplitudes by

$$V_n = \sqrt{K_n}\,(A_n + B_n),$$
$$I_n = \frac{1}{\sqrt{K_n}}\,(A_n - B_n), \tag{51}$$

where K_n is the wave impedance:

$$K_{(n)} = \frac{h_{(n)}}{\omega\epsilon}, \qquad K_{[n]} = \frac{\omega\mu}{h_{[n]}}. \tag{52}$$

If the currents and voltages in the generalized telegraphist's equations

are represented in terms of the traveling wave amplitudes, the system of coupled equations can be written in matrix notation as

$$\frac{dA}{dz} = -K'A. \tag{53}$$

The column matrix

$$A = \begin{bmatrix} A_{[0]} \\ A_{[1]} \\ A_{(1)} \\ A_{[2]} \\ A_{(2)} \\ \vdots \\ B_{[n]} \\ \vdots \end{bmatrix} \tag{54}$$

represents the amplitudes of metallic waveguide modes. The square matrix

$$K' = \begin{bmatrix} \kappa_{00} & \kappa_{01} & \kappa_{02} & \cdots \\ \kappa_{01} & \kappa_{11} & \kappa_{12} & \cdots \\ \kappa_{02} & \kappa_{12} & \kappa_{22} & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix} \tag{55}$$

describes the deformed helix waveguide. To calculate the elements of K' the wave functions of normal modes of metallic waveguide are introduced. The customary double-subscript notation will be used, but TM-waves will still be denoted with parentheses and TE-waves with brackets:

$$\begin{aligned} T_{(pn)} &= \sqrt{\frac{\epsilon_p}{\pi}} \frac{J_p(\chi_{(pn)}r) \sin p\varphi}{k_{(pn)}J_{p-1}(k_{(pn)})}, \\ T_{[pn]} &= \sqrt{\frac{\epsilon_p}{\pi}} \frac{J_p(\chi_{[pn]}r) \cos p\varphi}{(k_{[pn]}^2 - p^2)^{\frac{1}{2}}J_p(k_{[pn]})}, \end{aligned} \tag{56}$$

where

$$\begin{aligned} k_{(pn)} &= \chi_{(pn)}a, \qquad J_p(k_{(pn)}) = 0, \\ k_{[pn]} &= \chi_{[pn]}a, \qquad J_p'(k_{[pn]}) = 0 \end{aligned}$$

and

$$\begin{aligned} \epsilon_p &= 1 \quad \text{if} \quad p = 0, \\ \epsilon_p &= 2 \quad \text{if} \quad p \neq 0. \end{aligned}$$

Let the first element of A be the amplitude $A_{[0m]}$ of a circular electric wave. The propagation constant of this wave is

$$\kappa_{00} = jh_{[0m]} = j\sqrt{\omega^2\mu\epsilon - \chi_{[0m]}^2}.$$

The elements of the first row and column,

$$\kappa_{0n} = \begin{cases} \kappa_{[0m][pn]} = j\dfrac{\sqrt{\epsilon_p}}{2\pi}\dfrac{k_{[0m]}k_{[pn]}^2}{a^2\sqrt{h_{[0m]}h_{[pn]}}\sqrt{k_{[pn]}^2 - p^2}}\displaystyle\int_0^{2\pi}\delta\cos p\varphi\, d\varphi, \\ \kappa_{[0m](ps)} = 0, \end{cases} \tag{57}$$

describe the coupling between the circular electric wave and other modes as it is caused by the deformation of the helix waveguide. The κ_{0n} are independent of the wall impedance and the same as in metallic waveguide.

The other diagonal elements of K' are propagation constants of other modes:

$$\kappa_{nn} = \begin{cases} \kappa_{(pn)(pn)} = \gamma_{(pn)} = jh_{(pn)} + \dfrac{\epsilon_p}{2a}\dfrac{Z}{K_{(pn)}}, \\ \kappa_{[pn][pn]} = \gamma_{[pn]} = jh_{[pn]} + \dfrac{\epsilon_p}{2a}\dfrac{p^2}{k_{[pn]}^2 - p^2}\dfrac{Z}{K_{[pn]}}. \end{cases} \tag{58}$$

The off-diagonal elements describe coupling between the other modes:

$$\kappa_{nm} = \begin{cases} \kappa_{(pn)(pm)} = \dfrac{\epsilon_p}{2a}\dfrac{Z}{\sqrt{K_{(pn)}K_{(pm)}}}, \\ \kappa_{(pn)[pm]} = \kappa_{[pm](pn)} = -\dfrac{\epsilon_p}{2a}\dfrac{p}{\sqrt{k_{[pm]}^2 - p^2}}\dfrac{Z}{\sqrt{K_{(pn)}K_{[pm]}}}, \\ \kappa_{[pn][pm]} = \dfrac{\epsilon_p}{2a}\dfrac{p^2}{\sqrt{k_{[pn]}^2 - p^2}\sqrt{k_{[pm]}^2 - p^2}}\dfrac{Z}{\sqrt{K_{[pn]}K_{[pm]}}}. \end{cases}$$

They all depend on the wall impedance of the helix waveguide.

REFERENCES

1. Morgan, S. P. and Young, J. A., Helix Waveguide, B.S.T.J., **35,** 1956, p. 1347.
2. Unger, H. G., Helix Waveguide Theory and Application, B.S.T.J., **37,** 1958, p. 1599.
3. Morgan, S. P., Mode Conversion Losses in Transmission of Circular Electric Waves through Slightly Noncylindrical Guides, J. Appl. Phys. **21,** 1950, p. 329.
4. Rowe, H. E. and Warters, W. D., Transmission Deviations in Waveguide Due to Mode Conversion: Theory and Experiment, Proc. I.E.E., **106,** Pt. B, Suppl. 13, 1959, p. 30.
5. Schelkunoff, S. A., Conversion of Maxwell's Equations into Generalized Telegraphist's Equations, B.S.T.J., **34,** 1955, p. 995.
6. Rowe, H. E., to be published.
7. Unger, H. G., Normal Modes and Mode Conversion in Helix Waveguide, B.S.T.J., **40,** 1961, p. 255.
8. Unger, H. G., Noncylindrical Helix Waveguide, B.S.T.J., **40,** 1961, p. 233.

Winding Tolerances in Helix Waveguide

By H. G. UNGER

(Manuscript received September 29, 1960)

In a perfect helix waveguide the circular electric wave loss is increased by eddy currents, finite pitch of the helix, radiation through the wire spacing and effects of the wire coating. Only the contributions from eddy currents and pitch are large enough to limit wire size and spacing.

Experimental helix waveguides have tilted turns. These tilts cause coupling between circular electric and unwanted modes. From the coupling between modes in curved and in offset helix waveguide, the coupling in a helix waveguide with tilted turns is found. For helix waveguide with slightly irregular winding of arbitrary form, generalized telegraphist's equations are derived.

Tilts and other irregularities in the winding increase the circular electric wave loss. The average increase is a function of the covariance of irregularities. Winding tilts with an exponential covariance and an rms value of 0.6° increase the TE_{01} loss in 2-inch inside diameter waveguide at 55 kmc at the most by 10 per cent of the loss in a perfect copper pipe with smooth walls. Present fabrication procedures insure a smaller wire tilt than this.

I. INTRODUCTION

Helix waveguide consisting of closely wound insulated copper wire covered with an electrically absorbing or reactive jacket is a good transmission medium for circular electric waves.[1] In long distance communication with circular electric waves it is useful as a mode filter, for negotiating bends or particularly as transmission line proper instead of a plain metallic waveguide.

As in metallic waveguide, the loss of circular electric waves decreases steadily with frequency only in a perfect helix waveguide. Any deviations from a round and straight guide and from a uniform and low pitch will add to the loss of circular electric waves.

Deviations from straightness and deformations of the cross section of helix waveguide have been analyzed before and their effect on circular electric wave transmission has been determined.[2,3] When these imperfec-

tions are caused in the manufacturing process they are statistically distributed over the guide length with a small correlation distance. Then they add nearly the same average loss to the circular electric wave as they do in a plain metallic waveguide.[4] Manufacturing tolerances for straightness and for cross-sectional deformations are therefore the same for helix waveguide as they are for metallic waveguide.

Deviations of the winding from a low-pitch uniform spiral are imperfections peculiar to the helix waveguide. Their effect on circular electric wave transmission will be analyzed here and tolerances on the winding of the helix waveguide for low-loss transmission will be determined.

II. WIRE SIZE AND PITCH

This section reviews results of earlier work.

Helix waveguide is usually wound from round wire with an insulating layer. Even when such a helix is perfectly accurate and uniform its differences from a smooth metallic waveguide add to the circular electric wave loss. The various effects can be listed as follows:

2.1 *Eddy Current Losses in the Spaced Wires*[5,6]

The circumferential wall currents of circular electric waves are uniformly distributed in a smooth wall. In the spaced wires of the helix waveguide their distribution is nonuniform. The heat loss is therefore increased over the smooth wall loss. In Fig. 1 this loss increase is plotted over the spacing for a wire size small compared to the wavelength, using Morrison's calculations.

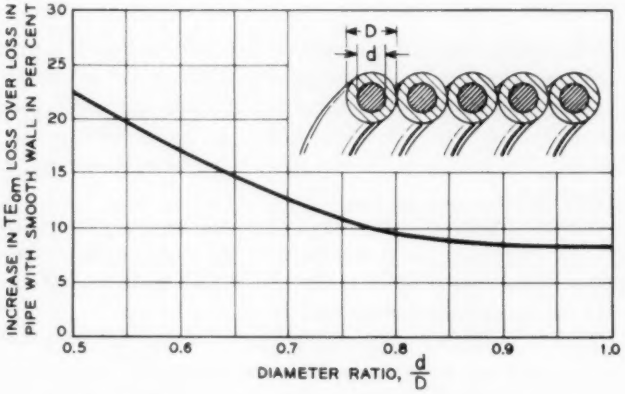


Fig. 1 — Eddy current losses in spaced helix wires (Ref. 5).

2.2 *Pitch*[1]

If the helix of radius a is closely wound from a single wire of insulation diameter D, then the pitch angle ψ is given by

$$\tan \psi = \frac{D}{2\pi a}.$$

If for a faster manufacturing process n wires are wound simultaneously,

$$\tan \psi = \frac{nD}{2\pi a}. \tag{1}$$

The wall currents of circular electric waves are strictly circumferential. In the helix waveguide their path is disturbed by the finite pitch. Power of circular electric waves is dissipated into the wall impedance Z, which the surrounding jacket presents to the waveguide interior. The added circular electric wave loss due to finite pitch is

$$\alpha_p = \frac{k_m^{\,2}}{kh_m a^3} \operatorname{Re}\left(\frac{Z}{Z_0}\right) \sin^2 \psi, \tag{2}$$

where $Z_0 = \sqrt{\mu/\epsilon}$ is the wave impedance and $k = \omega\sqrt{\mu\epsilon}$ the propagation constant of free space. k_m is the mth root of $J_1(x) = 0$, and $h_m = \sqrt{k^2 - (k_m^2/a^2)}$ the phase constant of the TE_{0m} wave.

A reactive jacket will not dissipate any power. A helix waveguide, designed for transmitting the circular electric wave around bends, has a quarter wave jacket with a very large wall impedance.[2] In this case, to keep α_p low, ψ has to be chosen small.

2.3 *Power Dissipation Through Wire Spacing*

Even though the helix is closely wound the wires are spaced by the wire insulation. With the electric field of circular electric waves parallel to the wires, the space between acts as waveguide below cutoff. Being short, this cutoff waveguide will transmit some circular electric wave power, which is then absorbed by the jacket. The circular electric wave loss caused by this power absorption has been investigated for various forms of wire cross section.[7] For round helix wires this loss is so small compared to the eddy current losses of Fig. 1 that it may be entirely neglected for any wire spacing. Consequently the increase in eddy current losses, rather than the power dissipation through the gaps, limits the wire spacing.

2.4 *Effect of Wire Insulation*

The insulating layer of the helix wires adds to the circular electric wave loss in two different ways. Its dielectric constant tends to concen-

trate the electric field into the layer. Thus the wall currents and the wall current losses are increased. In addition, the finite loss factor of any insulating material causes dielectric losses in the small but finite electric field of the circular electric wave. Both of these effects can be calculated with a sufficient approximation from attenuation formulas for the round waveguide with a dielectric lining.[8]

Again, it is found that the effects of the insulating layer are so small compared to the eddy current losses that they may be neglected.

Number of wires, wire size, and wire spacing through insulation are therefore determined by the pitch effect of (2) and the eddy current loss of Fig. 1. To speed the winding the numbers of wires should be large. To increase the effects of a reactive or resistive jacket on unwanted modes the wires should be widely spaced.[2] The increase in circular electric wave loss from Fig. 1 and equation (2) sets a limit, however, to number of wires and their spacing.

III. TILTED WINDING

The preceding discussion has considered only the loss in a perfectly wound helix waveguide. A practical helix will not have perfectly uniform windings. One imperfection in particular has been most notable in research models of helix waveguide made by Bell Telephone Laboratories. This imperfection is tilts in the winding.

Aside from the finite pitch, a single turn of the helix is usually not in a transverse plane, but is slightly inclined and forms a small angle θ with the axis. Even an improved winding method with an automatic feed control has not entirely eliminated this inclination.[9]

Such inclined helix turns give rise to mode conversion. There is a simple way to analyze circular electric wave propagation in helix waveguide with nonuniformly tilted winding, in which the results of previous calculations are used. Consider a perfect helix waveguide, a section of which between $z = 0$ and $z = L$ has been deflected in an arbitrary manner by $x(z)$, as shown in Fig. 2. With this deflection is associated a change of guide direction dx/dz and for gentle deflections a curvature $1/R = d^2x/dz^2$.

One way to calculate propagation through this deflected section is to use the formulas for wave propagation in the curved helix waveguide[2] and evaluate them for the curvature distribution d^2x/dz^2. Thus, when propagation is described by generalized telegraphist's equations,[10] there is curvature coupling between a circular electric mode m and the modes n

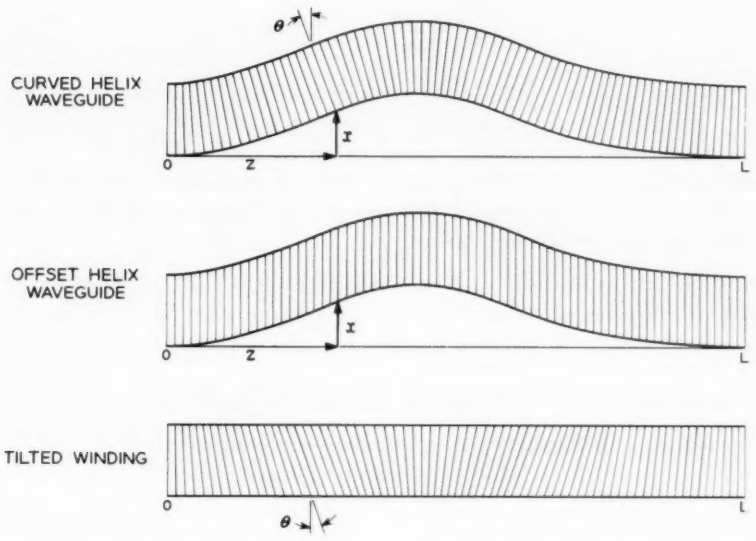


Fig. 2 — Curved helix waveguide as superposition of offset and tilted winding.

of first order circumferential dependence (i.e., TE_{1n} and TM_{1n}). The coupling coefficient is

$$c_c = C_c \frac{d^2x}{dz^2}, \tag{3}$$

with

$$C_c = N_n \frac{\sqrt{\pi}}{2ka} \sqrt{\frac{h_n}{h_m}} \frac{k_m k_n^2}{k_m^2 - k_n^2} J_1(k_n) \left(1 + \frac{h_m}{h_n} + \frac{h_m + h_n}{h_m - h_n} d_n\right) \tag{4}$$

where k, k_m and h_m have the same meanings as in (2). k_n is the radial propagation constant of the coupled mode n and $h_n = \sqrt{k^2 - (k_n^2/a^2)}$ is the axial propagation constant of the coupled mode n. N_n is a normalization factor for the coupled mode n [given by (35) below] and

$$d_n = \frac{pJ_p(k_n)}{k_n J_p'(k_n)}, \tag{5}$$

where $p = 1$ in the present case.

A mode m of unit amplitude incident at $z = 0$ converts power in the deflected section to the coupled modes n. For gentle deflection the amplitude and phase of mode n at $z = L$ is given by

$$A_n = -je^{-jh_nL} \int_0^L C_c \frac{d^2x}{dz^2} e^{-j(h_m-h_n)z}\, dz. \tag{6}$$

Another way to calculate propagation through the deflected section is to consider it as a continuously offset waveguide with a continuously varying tilt θ of the winding. Both the offset x and the tilt $\theta = dx/dz$ will then cause coupling between a circular electric mode m and modes n of first-order circumferential dependence. The coupling coefficient for offset has been calculated[3] before:

$$c_0 = C_0 x, \tag{7}$$

with

$$C_0 = N_n \frac{\sqrt{\pi}}{2} \sqrt{\frac{h_n}{h_m}} \frac{k_m k_n^2}{ka^3} J_1(k_n)\, d_n. \tag{8}$$

The coupling coefficient for tilted winding is still unknown:

$$c_t = C_t \frac{dx}{dz}. \tag{9}$$

Amplitude and phase of a coupled mode n are now given by

$$A_n = -je^{-jh_nL} \int_0^L \left(C_0 x + C_t \frac{dx}{dz} \right) e^{-j(h_m-h_n)z}\, dz. \tag{10}$$

Equation (10) should give the same result as (6). Integrating by parts brings (10) into a form that can be directly compared with (6):

$$A_n = -je^{-jh_nL} \int_0^L - \left(\frac{C_0}{(h_m - h_n)^2} + j \frac{C_t}{h_m - h_n} \right) \frac{d^2x}{dz^2} e^{-j(h_m-h_n)}\, dz. \tag{11}$$

Equations (11) and (6) can only given identical results when

$$C_c = -\frac{C_0}{(h_m - h_n)^2} - j \frac{C_t}{h_m - h_n}.$$

Hence the coupling coefficient for tilted winding is

$$jC_t = (h_n - h_m)C_c + \frac{C_0}{h_n - h_m}. \tag{12}$$

Substituting from (4) and (8) into (12), and from (12) into (9):

$$jc_t = N_n \frac{\sqrt{\pi}}{2} \frac{k_m k_n^2}{k\sqrt{h_m h_n}\, a^3} J_1(k_n)\theta. \tag{13}$$

With these coupling coefficients, generalized telegraphist's equations

can be written down for the helix waveguide with tilted winding. Their solution describes propagation in the obliquely wound helix waveguide completely. For example, if a mode m of unit amplitude is incident on a length L of helix waveguide with nonuniformly tilted winding $\theta(z)$, then the output amplitude and phase of this mode m is given to second order by

$$A_m = e^{-jh_m L}\left[1 - \sum_n \int_0^L e^{j(h_m - h_n)z}\, dz \int_0^{L-z} c_{tn}(u)c_{tn}(u+z)\, du\right]. \quad (14)$$

The summation in (14) is to be extended not only over all the modes n but also over their two polarizations according to the orientation of θ; θ may not only be in the plane of Fig. 2, it can also be perpendicular to that plane.

From (14) the loss which is added to the mode m by a tilted winding may be calculated. With

$$j(h_m - h_n) = \Delta\alpha_n + j\Delta\beta_n$$

and

$$C_{tn}^2 = P_n + jQ_n\,,$$

the added loss is to second order in θ is

$$\alpha_t = \frac{1}{L}\sum_n \int_0^L e^{\Delta\alpha_n z}(P_n \cos\Delta\beta_n z - Q_n \sin\Delta\beta_n z)\, dz \cdot \int_0^{L-z} \theta(u)\theta(u+z)\, du. \quad (15)$$

IV. IRREGULAR WINDING

The obliquely wound helix of the preceding section is just a special case of a general irregular winding. In Fig. 3 a turn of such an oblique helix has been drawn in more detail. Aside from a small pitch the wire follows the curve

$$z = a \tan\theta(1 - \cos\varphi) \quad (16)$$

around the circumference. Its direction deviates by ψ from the transverse direction, where

$$\tan\psi = \tan\theta \sin\varphi$$

or, to first order for small tilt,

$$\psi = \theta \sin\varphi. \quad (17)$$

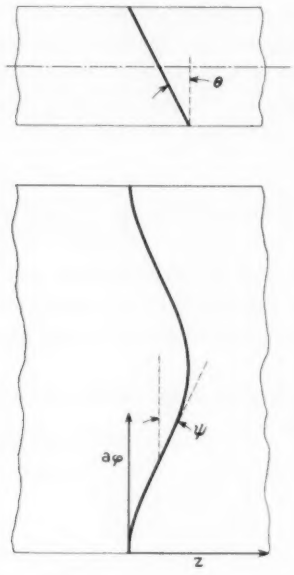


Fig. 3 — Tilted turn in helix waveguide.

A general irregular winding can be described by a Fourier series

$$\psi = \sum_p \theta_p \sin p\varphi. \tag{18}$$

A summation of cos $p\varphi$ would only add identical terms with different polarization; it has been omitted from (18). The boundary conditions at this irregular helix are

$$E_\varphi = -E_z \tan \psi, \tag{19}$$

$$E_z = -ZH_\varphi + (E_\varphi - ZH_z) \tan \psi, \tag{20}$$

where Z is the wall impedance which the outside jacket presents through the helix to the waveguide interior.

Wave propagation in such an irregular structure is best analyzed by converting Maxwell's equations:

$$\frac{1}{r}\frac{\partial E_z}{\partial \varphi} - \frac{\partial E_\varphi}{\partial z} = -j\omega\mu H_r, \tag{21}$$

$$\frac{\partial E_r}{\partial z} - \frac{\partial E_z}{\partial r} = -j\omega\mu H_\varphi, \tag{22}$$

$$\frac{1}{r}\frac{\partial}{\partial r}(rE_\varphi) - \frac{1}{r}\frac{\partial E_r}{\partial \varphi} = -j\omega\mu H_z \tag{23}$$

$$\frac{1}{r}\frac{\partial H_z}{\partial \varphi} - \frac{\partial E_\varphi}{\partial z} = j\omega\epsilon E_r\,, \tag{24}$$

$$\frac{\partial H_r}{\partial z} - \frac{\partial H_z}{\partial r} = j\omega\epsilon E_\varphi\,, \tag{25}$$

$$\frac{1}{r}\frac{\partial}{\partial r}(rH_\varphi) - \frac{1}{r}\frac{\partial H_r}{\partial \varphi} = j\omega\epsilon E_z \tag{26}$$

into generalized telegraphist's equations[10] for the boundary conditions (19) and (20).

An appropriate form of representation is in terms of normal modes of the perfect helix waveguide. With two sets of wave functions

$$\begin{aligned} T_n &= N_n J_p(\chi_n r)\sin p\varphi, \\ T_n' &= N_n J_p(\chi_n r)\cos p\varphi \end{aligned} \tag{27}$$

which satisfy the wave equation

$$\nabla^2 T_n = -\chi_n{}^2 T_n \tag{28}$$

the normal mode fields of the perfect helix waveguide are the individual terms of the sums:

$$\begin{aligned} E_r &= \sum_n V_n\left(\frac{\partial T_n}{\partial r} + \frac{d_n}{r}\frac{\partial T_n'}{\partial \varphi}\right), \\ E_\varphi &= \sum_n V_n\left(\frac{\partial T_n}{r\partial \varphi} - d_n\frac{\partial T_n'}{\partial r}\right), \\ H_r &= \sum_n I_n\left(-\frac{\partial T_n}{r\partial \varphi} + d_n\frac{h_n{}^2}{k^2}\frac{\partial T_n'}{\partial r}\right), \\ H_\varphi &= \sum_n I_n\left(\frac{\partial T_n}{\partial r} + \frac{d_n}{r}\frac{h_n{}^2}{k^2}\frac{\partial T_n'}{\partial \varphi}\right), \end{aligned} \tag{29}$$

and by substituting from (29) into (23) and (26) and using (28):

$$\begin{aligned} H_z &= j\omega\epsilon\sum_n V_n d_n\frac{\chi_n{}^2}{k^2}T_n', \\ E_z &= j\omega\mu\sum_n I_n\frac{\chi_n{}^2}{k^2}T_n\,. \end{aligned} \tag{30}$$

From $E_\varphi = 0$ at the perfect helix:

$$d_n = \frac{p}{k_n} \frac{J_p(k_n)}{J_p'(k_n)}, \tag{31}$$

and from $E_z = -ZH_\varphi$ again for the perfect helix:

$$\frac{p}{k_n^2} \left(\frac{h_n^2}{k^2} d_n - \frac{1}{d_n} \right) = \frac{j}{\omega \epsilon a Z}. \tag{32}$$

Equation (32) is the characteristic equation for the perfect helix waveguide. Its roots $k_n = \chi_n a$ determine the propagation constants

$$h_n^2 = k^2 - \chi_n^2 \tag{33}$$

of the normal modes of the perfect helix waveguide.

The transverse fields of the normal modes are orthonormal in that

$$\frac{1}{V_n I_m} \int_S (E_{tn} \times H_{tm}) \, dS = \delta_{nm} \tag{34}$$

when the normalization factor N_n in (27) is chosen so that

$$N_n = \frac{\sqrt{2}}{\sqrt{\pi} J_p(k_n)} \left[\frac{h_n^2}{k^2} (k_n^2 - p^2) d_n^2 + \frac{p^2}{d_n^2} + k_n^2 \left(1 - \frac{p^2}{k^2 a^2} \right) + 2p \left(\frac{1}{d_n} - d_n \right) \right]^{-\frac{1}{2}}. \tag{35}$$

The integral in (34) extends over the cross section of the guide; δ_{nm} is the Kronecker symbol.

The z-dependence of the voltage and current coefficients in (29) is found by substituting the sums of (29) for the transverse field components into Maxwell's equations.

Add

$$-\left(\frac{1}{r} \frac{\partial T_m}{\partial \varphi} - d_m \frac{h_m^2}{k^2} \frac{\partial T_m'}{\partial r} \right)$$

times (21) and

$$\frac{\partial T_m}{\partial r} + d_m \frac{h_m^2}{k^2} \frac{1}{r} \frac{\partial T_m'}{\partial \varphi}$$

times (22) and integrate over the cross section of the guide. Using (34), the result is

$$\frac{dV_m}{dz} + j \frac{h_m^2}{\omega \epsilon} I_m = \int_S (\text{grad } E_z)(\text{grad } T_m) \, dS + d_m \frac{h_m^2}{k^2} \int_S (\text{grad } E_z)(\text{flux } T_m') \, dS - j\omega\mu \sum_n I_n \frac{\chi_n^2}{k^2} \cdot \int_S \left[(\text{grad } T_n)(\text{grad } T_m) + d_m \frac{h_m^2}{k^2} (\text{grad } T_n)(\text{flux } T_m') \right] dS, \tag{36}$$

where the gradient and flux of a scalar are defined by:

$$\begin{aligned} \text{grad}_r\, T &= \frac{\partial T}{\partial r}, \qquad \text{grad}_\varphi\, T = \frac{1}{r}\frac{\partial T}{\partial \varphi}, \\ \text{flux}_r\, T &= \frac{1}{r}\frac{\partial T}{\partial \varphi}, \qquad \text{flux}_\varphi\, T = -\frac{\partial T}{\partial r}. \end{aligned} \tag{37}$$

After partial integration on the right-hand side of (36),

$$\begin{aligned} \frac{dV_m}{dz} + j\frac{h_m^2}{\omega\epsilon} I_m &= \int_0^{2\pi} E_z \left(\frac{\partial T_m}{\partial r} + \frac{d_m}{a}\frac{h_m^2}{k^2}\frac{\partial T_m'}{\partial \varphi}\right) a\, d\varphi \\ &+ \chi_m^2 \int_S E_z T_m\, dS - j\omega\mu \sum_n I_n \frac{\chi_n^2}{k^2} \\ &\cdot \left[\int_0^{2\pi} T_n \left(\frac{\partial T_m}{\partial r} + \frac{d_m}{a}\frac{h_m^2}{k^2}\frac{\partial T_m'}{\partial \varphi}\right) a\, d\varphi + \chi_m^2 \int_S T_n T_m\, dS\right]. \end{aligned} \tag{38}$$

In the line integral along the boundary, E_z from the boundary condition (20) is substituted. In the surface integral over the cross section, (30) is substituted for E_z. Subsequently the boundary conditions of the perfect helix waveguide may be used to simplify (38) to first order in ψ:

$$\frac{dV_m}{dz} + j\frac{h_m^2}{\omega\epsilon} I_m = -\sum_n V_n d_n \frac{k_n^2 k_m^2}{k^2 a^3} \int_0^{2\pi} T_n' T_m \psi\, d\varphi. \tag{39}$$

For the other set of generalized telegraphist's equations add

$$-\left(\frac{\partial T_m}{\partial r} + \frac{d_m}{r}\frac{\partial T_m'}{\partial \varphi}\right)$$

times (24) and

$$-\left(\frac{1}{r}\frac{\partial T_m}{\partial \varphi} - d_m \frac{\partial T_m'}{\partial r}\right)$$

times (25) and integrate over the cross section. The result is

$$\begin{aligned} \frac{dI_m}{dz} + j\omega\epsilon V_m &= -\int_S (\text{grad}\, H_z)(\text{flux}\, T_m)\, dS \\ &+ d_m \int_S (\text{grad}\, H_z)(\text{grad}\, T_m')\, dS + j\omega\epsilon \sum_n V_n d_n \frac{\chi_n^2}{k^2} \\ &\cdot \int_S [(\text{grad}\, T_n')(\text{flux}\, T_m) - d_m (\text{grad}\, T_n')(\text{grad}\, T_m')]\, dS. \end{aligned} \tag{40}$$

After partial integration on the right-hand side of (40):

$$\frac{dI_m}{dz} + j\omega\epsilon V_m =$$

$$d_m\chi_m^2 \int_S H_z T_m' \, dS - j\omega\epsilon \sum_n V_n d_n d_m \frac{\chi_n^2 \chi_m^2}{k^2} \int_S T_n' T_m' \, dS. \tag{41}$$

Expression (30) for H_z holds only for the normal modes of the perfect waveguide. It has been obtained by differentiating the sum (29) for E_φ in (23). The individual terms of this sum vanish at $r = a$, while in the present case according to (19) E_φ has a finite value there. Hence the sum (29) for E_φ is nonuniformly convergent and differentiation makes it diverge. To replace H_z in (41), substitute in (23) E_r from (29), multiply (23) by T_m', and integrate over the cross section

$$-j\omega\mu \int_S H_z T_m' \, dS = \int_0^{2\pi} E_\varphi T_m' a \, d\varphi + \sum_n V_n d_n \chi_n^2 \int_S T_n' T_m' \, dS. \tag{42}$$

With (42), the boundary condition (19) for E_φ, and (30) for E_z, the second set of generalized telegraphist's equations is

$$\frac{dI_m}{dz} + j\omega\epsilon V_m = \sum_n I_n d_m \frac{k_n^2 k_m^2}{k^2 a^3} \int_0^{2\pi} T_n T_m' \psi \, d\varphi. \tag{43}$$

These equations represent an infinite set of coupled transmission lines. For the present purpose it is more convenient to write these transmision-line equations not in terms of currents and voltages but in terms of the amplitudes of forward and backward traveling waves A and B. The current and voltage of a typical mode are related to the wave amplitudes by

$$V = \sqrt{K}\,(A + B),$$

$$I = \frac{1}{\sqrt{K}}\,(A - B),$$

where K is the wave impedance

$$K_m = \frac{h_m}{\omega\epsilon}.$$

If the currents and voltages in (39) and (43) are replaced by the traveling wave amplitudes, the following equations for coupled traveling waves are obtained:

$$\begin{aligned} \frac{dA_m}{dz} + jh_m A_m &= \sum_n (\kappa_{nm}^{+} A_n - \kappa_{nm}^{-} B_n), \\ \frac{dB_m}{dz} - jh_m B_m &= \sum_n (\kappa_{nm}^{+} B_n - \kappa_{nm}^{-} A_n). \end{aligned} \tag{44}$$

The coupling coefficients are given by

$$\kappa_{nm}^{\pm} = \frac{k_n^2 k_m^2}{2k^2 a^3}\left(d_m\sqrt{\frac{h_m}{h_n}}\int_0^{2\pi} T_n T_m'\psi\, d\varphi \mp d_n\sqrt{\frac{h_n}{h_m}}\int_0^{2\pi} T_n' T_m\psi\, d\varphi\right). \tag{45}$$

If m represents a circular electric wave:

$$T_m = 0,$$

$$T_m' = N_m J_0(\chi_m r),$$

$$d_m N_m = \frac{k}{\sqrt{\pi}\, h_m k_m J_0(k_m)},$$

and for the coupling coefficients,

$$\kappa_{nm}^{\pm} = \frac{k_m k_n^2}{2\sqrt{\pi}\, k\sqrt{h_m h_n}\, a^3}\int_0^{2\pi} T_n\psi\, d\varphi. \tag{46}$$

For a tilted winding with ψ from (17),

$$\kappa_{nm}^{\pm} \equiv jc_t.$$

In this case circular electric waves interact only with modes of first circumferential order ($p = 1$). Helix irregularities of higher order in ψ will cause coupling to modes of correspondingly higher circumferential order. In general,

$$\kappa_{nm}^{\pm} = \frac{\sqrt{\pi}\, k_m k_n^2 N_n J_p(k_n)}{2k\sqrt{h_m h_n}\, a^3}\theta_p. \tag{47}$$

V. TOLERANCES

The design of a helix waveguide is started by selecting wire size and spacing. A tolerable amount of added TE_{01} loss is specified, and with Fig. 1 and equation (2) wire size and spacing are determined.

If, for example, eddy current losses in the helix should not be more than 10 per cent of the loss in a guide with smooth walls, then from Fig. 1 the ratio of wire diameter to insulation diameter should be

$$\frac{d}{D} > 0.775. \tag{48}$$

To determine the actual wire size with (2) the wall impedance has to be specified. Different applications of the helix waveguide require different values for the wall impedance. A typical and also very critical ex-

ample is the helix waveguide for intentional bends. In this case, by surrounding the helix with a quarter-wave jacket and a metallic shield the wall impedance is made very high.

The general formula for the wall impedance of the shielded helix waveguide is[2]

$$Z = j \frac{k_n^e}{\omega \epsilon_e a} \tan k_n^e \delta, \tag{49}$$

where ϵ_e is the permittivity of the jacket, $k_n^e = a\sqrt{\omega^2 \mu \epsilon_e - h_n^2}$ the radial propagation constant in it, and δ the relative thickness.

For a quarter-wave jacket of a low-loss material the wall impedance is real and approximately

$$\frac{Z}{Z_0} = \frac{4}{\pi} \frac{(\epsilon' - 1)^{\frac{1}{2}}}{\epsilon' \epsilon''}, \tag{50}$$

with

$$\frac{\epsilon_e}{\epsilon_0} = \epsilon' - j\epsilon''.$$

Substituting (50) into (2), an equation for nD is obtained.

Fiber glass laminated with epoxy resin has a relative permittivity at millimeter wavelengths of $\epsilon_e/\epsilon_0 = 4 - j(0.04)$. The relative wall impedance from (49) is then $Z/Z_0 = 41.4$. In 2-inch inside diameter waveguide with smooth walls, the TE_{01} loss at 55.5 kmc is $\alpha_0 a = 2.77 \times 10^{-6}$. Less than 10 per cent of this figure is added to the TE_{01} loss in the present example when the pitch is

$$\frac{nD}{a} < 4 \times 10^{-3}. \tag{51}$$

No. 37 wire (AWG) with a heavy Formex coat has $d = 0.0045$ and $D = 0.0054$. It very nearly satisfies conditions (48) and (51) when the helix is wound from one wire only ($n = 1$). Lower wall impedance values such as are used for helix mode filters or all helix guide would not require as low a pitch as (51).

For the winding process, tolerances for irregularities must be specified. In (15) the added loss is expressed in terms of the θ of a tilted winding. The loss caused by higher-order irregularities can also be determined by (15) when the corresponding coupling coefficients (47) are substituted.

In the present problem, however, the irregularities are not known, but at best some of their statistical properties are known. Equation (15) can then be used to express the statistics of the loss in terms of the statistics of the winding irregularities.[11,12]

For an oblique winding θ is assumed to be a stationary random process with covariance $R(u)$ and spectral distribution $S(\xi)$:

$$R(u) = \langle\theta(z)\theta(z + u)\rangle, \tag{52}$$

$$S(\xi) = \int_{-\infty}^{+\infty} R(u)\, e^{-j2\pi\xi u}\, du. \tag{53}$$

In (52), $\langle x\rangle$ is the expected value of x.

Taking the expected value on both sides of (15) the average added loss is obtained in terms of the covariance $R(u)$:

$$\langle\alpha_t\rangle = \frac{1}{L}\sum_n \int_0^L e^{\Delta\alpha_n z} R(z)(L - z)(P_n \cos\Delta\beta_n z - Q_n \sin\Delta\beta_n z)\, dz. \tag{54}$$

For a mere estimate the covariance is assumed to be exponential to simplify the calculation:

$$R(z) = \frac{\pi S_0}{L_0}\, e^{(-2\pi|z|/L_0)}. \tag{55}$$

Then the spectral distribution of θ becomes

$$S(\xi) = \frac{S_0}{1 + L_0^2\xi^2}, \tag{56}$$

with $S(\xi)$ nearly flat with spectral density S_0 for mechanical frequencies smaller than $\xi_0 = 1/L_0$. L_0 may be regarded as the cutoff mechanical wavelength according to (56) or as a correlation distance according to (55).

The average added loss is for $L \gg L_0$

$$\langle\alpha_t\rangle = \pi S_0 \sum_n \frac{P_n(2\pi - \Delta\alpha_n L_0) - Q_n\Delta\beta_n L_0}{\Delta\beta_n^2 L_0^2 + (2\pi - \Delta\alpha_n L_0)^2}. \tag{57}$$

To evaluate (57) the characteristic equation (32) of the perfect helix waveguide has to be solved for all the coupled modes n, and propagation constants and coupling coefficients have to be calculated.

The helix waveguide for intentional bends is again a typical and critical example. In this case $Z = \infty$ at the design frequency and the characteristic equation (32) simplifies to

$$d_n = \pm\frac{k}{h_n}. \tag{58}$$

The roots k_{n0} of $J_{p+1}(x) = 0$ and $J_{p-1}(x) = 0$ are good approximations for the roots of (58). With $k_n = k_{n0} \pm x$, where

$$x = \frac{p}{k_{n0}}\left(1 - \sqrt{1 - \frac{k_{n0}}{k^2 a^2}}\right), \tag{59}$$

the approximations can be sufficiently improved. The coupling coefficients in this case are given by

$$\kappa_{nm} = \frac{k_m k_n \theta_p}{2k\sqrt{h_m h_n}\, a^3}\left(1 - \frac{p^2}{k^2a^2} \mp \frac{p}{kh_n a^2}\right)^{-\frac{1}{2}}. \tag{60}$$

With these relations, (57) has been evaluated for a helix waveguide with a nonuniformly tilted winding ($p = 1$).

Fig. 4 shows the rms value of θ as a function of the correlation distance L_0 for an added average TE_{01} loss of 10 per cent of the loss in a copper pipe with smooth walls. The waveguide diameter is 2 inches and the frequency 55.5 kmc. The tolerance is most critical for a correlation distance of 1 inch. But even then an rms tilt of 0.6° can be tolerated. In experimental models of helix waveguide the maximum tilt has, with some care, been kept below 0.3°.

VI. CONCLUSION

In a perfectly wound helix waveguide the circular electric wave loss is significantly increased only by the eddy current losses in spaced wires. The finite pitch contributes to the circular electric wave loss only when the wall impedance is very high or when the helix is wound from more than one wire.

Of all irregularities in the helix a changing tilt of the winding has been observed to be most significant. Assuming this tilt to be randomly

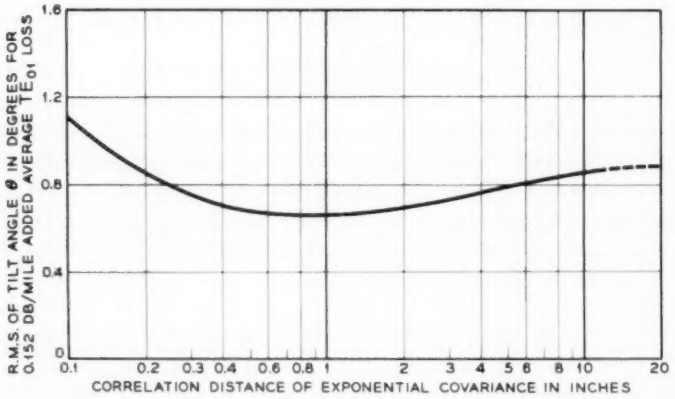


Fig 4. — TE_{01} loss in helix waveguide with random tilt of winding, 2-inch inside diameter, at 55.5 kmc. Design for intentional bends with infinite wall impedance.

distributed along the waveguide with an exponential covariance, the increase in circular electric wave loss can be calculated. In a 2-inch inside diameter helix waveguide at 55.5 kmc an rms tilt angle of 0.6° adds at the most 10 per cent of the loss in a perfect copper pipe to the average TE_{01} loss. In experimental models of helix waveguide the maximum tilt deviation has been kept below 0.3°.

REFERENCES

1. Morgan, S. P., and Young, J. A., Helix Waveguide, B.S.T.J., **35,** 1956, p. 1347.
2. Unger, H. G., Helix Waveguide Theory and Application, B.S.T.J., **37,** 1958, p. 1599.
3. Unger, H. G., Noncylindrical Helix Waveguide, B.S.T.J., **40,** 1961, p. 233.
4. Unger, H. G., Mode Conversion in Metallic and Helix Waveguide, this issue, p. 613.
5. Marcatili, E. A., Heat Loss in Grooved Metallic Surface, Proc. I.R.E., **45,** 1957, p. 1134.
6. Morrison, J. A., Heat Loss of Circular Electric Waves in Helix Waveguides, I.R.E., Trans., **MTT-6,** 1958, p. 173.
7. Katsenelenbaum, B. Z., Attenuation of H_{on} Modes in a Helical Waveguide, Radiotekhnika i Elektronika **4,** 1959, p. 428.
8. Unger, H. G., Circular Electric Wave Transmission in a Dielectric-Coated Waveguide, B.S.T.J., **36,** 1957, p. 1253.
9. Beck, A. C. and Rose, C. F. P., Waveguide for Circular Electric Mode Transmission, Proc. I.E.E., **106,** Pt. B, Suppl. 13, 1959, p. 159.
10. Schelkunoff, S. A., Conversion of Maxwell's Equations into Generalized Telegraphist's Equations, B.S.T.J., **34,** 1955, p. 995.
11. Rowe, H. E. and Warters, W. D., Transmission Deviations in Waveguide Due to Mode Conversion: Theory and Experiment, Proc. I.E.E., **106,** Pt. B, Suppl. 13, 1959, p. 3.
12. Rowe, H. E., to be published.

Contributors to This Issue

D. B. Armstrong, B.A., 1940, University of Toronto; M.S., 1951, and Sc.D., 1955, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1954—. He has been engaged in research in switching problems, which have included simulation and economic studies of telephone systems, studies of central office systems and reliability studies of digital systems. Member Sigma Xi.

D. W. Bodle, B.S. in E.E., 1937, New York University; Bell Telephone Laboratories, 1935—. He has been concerned with problems of electric shock and the protection of communication facilities from foreign potentials. This has included field investigations of lightning behavior, design of surge measuring devices and laboratory surge testing of apparatus. Member A.I.E.E.

G. D. Boyd, B.S., 1954, M.S., 1955, and Ph.D., 1959, California Institute of Technology; Bell Telephone Laboratories, 1959—. He is engaged in optical maser research.

Donald B. Cuttriss, B.S. in E.E., 1959, Newark College of Engineering; Bell Telephone Laboratories, 1951—. Until 1959 Mr. Cuttriss was a member of the component development department working on design and development of voice-frequency laminated-core inductors and semiconductor field-effect devices. In 1959 he transferred to transistor development and he has been concerned with development of diffused-base germanium transistors and with methods of producing high-quality gallium arsenide. Member Tau Beta Pi.

A. Gardner Fox, S.B., 1934, and S.M., 1935, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1936—. His early work was in development of mobile radio transmitters, early radar development and general waveguide research. Since 1944 he has been engaged in design of radio frequency amplifiers and in research on millimeter waves. He heads a group specializing in microwave physics. Fellow I.R.E.

JAMES P. GORDON, B.S., 1949, Massachusetts Institute of Technology, M.A., 1951, and Ph.D., 1955, Columbia University; Bell Telephone Laboratories, 1955—. His research in quantum electronics has involved work on molecular beam masers, paramagnetic resonance and solid state masers. Member American Association for the Advancemens of Science, American Physical Society, Sigma Xi.

PHILIP A. GRESH, B.S.E.E., 1956, Carnegie Institute of Technology; Bell Telephone Laboratories, 1956—. His work in Systems Engineering has included statistical and economic optimization studies of outside plant facilities. Member Eta Kappa Nu, Phi Kappa Phi, Tau Beta Pi.

TINGYE LI, B.Sc., 1953, University of Witwatersrand (South Africa); M.S., 1955, and Ph.D., 1958, Northwestern University; Bell Telephone Laboratories, 1957—. He has been engaged in studies of microwave antennas and microwave propagation. Recently he has been primarily concerned with work on optical masers. Member I.R.E., Eta Kappa Nu, Sigma Xi.

H. C. MARTEL, B.S., 1949, and Ph.D., 1956, California Institute of Technology; M.S., 1950, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1959–60; associate professor of electrical engineering, California Institute of Technology, 1953—. While at Bell Laboratories during a year's leave of absence from Cal Tech, Mr. Martel was engaged in visual and acoustics research related to signal coding and detection. Member I.R.E., Sigma Xi, Tau Beta Pi.

MAX V. MATHEWS, B.S., 1950, California Institute of Technology; M.S., 1952, and Sc.D., 1954, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1955—. He has specialized in acoustics research in speech transmission and has been especially concerned with simulating speech experiments on a digital computer. Member Acoustical Society of America, I.R.E., Sigma Xi.

R. L. PEEK, JR., A.B., 1921, Columbia College; Met.E., 1923, Columbia School of Mines; Bell Telephone Laboratories, 1924—. His early work was in materials testing. In 1936 he turned to apparatus development involving coin collectors, electromagnets, relays and switches. During the war he was engaged in military work and after the war in wire-spring relay development. Since 1951 he has been in charge of a group studying new electromagnetic devices. Member A.I.E.E.

D. K. Ray-Chaudhuri, B.Sc., 1953, Presidency College, University of Calcutta (India); M.Sc., 1955, University of Calcutta; Ph.D., 1959, University of North Carolina; instructor in statistics, Case Institute of Technology, 1959–60; Bell Telephone Laboratories, summer, 1960; visiting assistant professor of statistics, University of North Carolina, 1960—. He has been engaged in research on problems of constructing experimental designs in statistics, sampling theory and error-correcting codes. Member Institute of Mathematical Statistics.

Erling D. Sunde, Dipl.Ing., 1926, Technische Hochschule, Darmstadt, Germany; American Telephone & Telegraph Co., 1927–34; Bell Telephone Laboratories, 1934—. He has made theoretical and experimental studies of inductive interference from railway and power systems, lightning protection of the telephone plant, and fundamental transmission studies in connection with the use of pulse modulation systems. Author of *Earth Conduction Effects in Transmission Systems*, a Bell Laboratories Series book. Senior member I.R.E.; member American Association for the Advancement of Science, A.I.E.E., American Mathematical Society.

Hans-Georg Unger, Dipl. Ing., 1951 and Dr. Ing., 1954, Technische Hochschule, Braunschweig (Germany); Siemens and Halske (Germany), 1951–55; Bell Telephone Laboratories, 1956—. His work at Bell Laboratories has been in research in waveguides, especially circular electric wave transmission. He is now on leave of absence from Bell Laboratories as professor of electrical engineering at the Technische Hochschule in Braunschweig. Senior member I.R.E.; member German Communication Engineering Society.